PRAVDEPODOBNOSŤ A MATEMATICKÁ ŠTATISTIKA

PETER VOLAUF

SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
2014

Učebný text sa zaoberá základmi teórie pravdepodobnosti a matematickej štatistiky. Teórii pravdepodobnosti je venovaných päť prvých kapitol a ostatné tri sú základné kapitoly matematickej štatistiky. Posledná kapitola, ktorá je venovaná štatistickým testom, je napísaná tak, aby poslúžila študentom aj pri ich ďalšej tvorivej práci v aplikovanej informatike.
Učebnica od študenta predpokladá zvládnutie základov diferenciálneho a integrálneho počtu aspoň pre funkcie jednej premennej. Je napísaná tak, aby teória neodrádzala, ale definície sú (pre väčšiu zrozumiteľnosť) formulované ako definície. Rovnako tvrdenia sú formulované ako matematické propozície (vety), pričom dôkazy sú uvedené len tam, kde napomáhajú pochopeniu súvislostí. Dôraz je venovaný motivačným a ilustračným príkladom. Učebnica ponúka študentovi pomerne veľké množstvo (viac ako 200) nevyriešených úloh (ale s uvedenými výsledkami), takže môže slúžiť aj ako zbierka úloh.

Publikáciu podporili ZDRUŽENIE GRATEX IT INŠTITÚT v rámci fondu GraFIIT www.gratex.com

Recenzenti: prof. RNDr. Radko Mesiar, DrSc.
doc. RNDr. Vladimír Olejček, PhD.

Schválilo Vedenie Fakulty informatiky a informačných systémov STU v Bratislave.

ISBN 978-80-227-4144-6

OBSAH

Predslov

Najprv smerom k študentom. Prosím, majme na zreteli, že zvládnuť pravdepodobnosť a štatistiku neznamená len oboznámiť sa s postupmi či schémami, ktoré ponúka teória, ale dokázať ich uplatniť v reálnej situácii. To vyžaduje vedieť vybrať správny model, poznať techniku riešenia a nakoniec korektne interpretovať získaný výsledok. Pre študenta niet inej cesty, ako prepočítať čo najviac (pokiaľ možno zaujímavých) úloh, pretože práve cez ich riešenie získa schopnosť poradiť si v reálnej situácii.

Nové pojmy a súvislosti sú v tomto texte objasňované najprv v rámci riešených príkladov (je ich naozaj dosť, viac ako 130). Definície a vety nastupujú až v druhom slede. Aj keď zďaleka nie všetky vety dokazujeme, vždy som sa snažil objasniť ich zmysel a ilustrovať ich použitie. Učebnica ponúka viac ako 200 neriešených úloh, takže môže slúžiť aj ako zbierka úloh. Aj keď som sa, pochopiteľne, snažil vyhnúť chybám, nemôžem s istotou povedať, že text už žiadne neobsahuje. Preto očakávam, že ak to bude potrebné, vytvoríme na Internete priestor, kde umiestnime opravenku (errata). Nakoniec, prajem čitateľom veľa chuti a energie pri „pasovačke" s PaS.

Teraz niekoľko myšlienok ku kolegom, ktorí budú pri svojej práci text používať. Úvodný kurz pravdepodobnosti a štatistiky by mal pripraviť študenta do tej miery, aby si v budúcich kontaktoch so stochastikou vedel poradiť. Pre študenta informatiky to napr. znamená byť pripravený študovať problematiku Markovových reťazcov. Aj na toto som myslel pri koncipovaní (pomerne obsiahlej) prvej kapitoly, konkrétne článku o podmienenej pravdepodobnosti. Na druhej strane, pri spolupráci s informatikmi som pochopil, že keď vytvoria nejaký IT produkt, potrebujú disponovať prostriedkami, ktorými preukážu prínos svojho výrobku (mám na mysli napr. štatistické testy o stredných hodnotách dvoch náhodných veličín). Aj keď to kolegyne Majka a Gabika nemôžu tušiť, práve oni iniciovali zaradenie článku 8.4 do tejto učebnice.

Pretože predmet PaS sa ocitol (na moje prekvapenie) už v prvom ročníku FIIT, zo štatistiky v ňom ide len o absolútne základy. Skutočná štatistika, samozrejme, pracuje s reálnymi dátami, čo dnes znamená pracovať v nejakom výpočtovom prostredí. To by sa mohlo realizovať, podľa mojej mienky, v rámci ďalšieho predmetu, tentoraz venovanému výhradne štatistike (lebo základy pravdepodobnosti podané v tomto texte iste postačujú).

Nakoniec by som rád poďakoval obom recenzentom za ochotu venovať sa tomuto textu, za cenné pripomienky, ktoré text zlepšili a za priaznivú kritiku. Moje poďakovanie patrí aj Katke Jelemenskej, ktorá – možno to nevie – veľkou mierou ovplyvnila moje učiteľovanie (vďaka akejsi zhode náhod a poskytnutí HP tabletu). Ďakujem aj Majke Bielikovej za dôveru a za čas, ktorý mi ochotne venovala pri finálnom formátovaní textu.

1 MATEMATICKÉ MODELY NÁHODNÝCH POKUSOV

1.1 Náhodné udalosti a pravdepodobnosť

Najprv objasnime, čo budeme rozumieť pod *náhodným pokusom*. Slovo *pokus* v tomto texte chápeme v širšom zmysle. Často to bude skutočne nejaký pokus, akým je napríklad sledovanie trajektórie pohybujúceho sa bodu po zadanej mriežke, ale niekedy pôjde jednoducho o situácie z bežného života. Adjektívum *náhodný* znamená, že výsledok pokusu nebude jednoznačne určený podmienkami pokusu. Ak ide o bežné situácie, tak prítomnosť náhodnosti chápeme najčastejšie ako dôsledok toho, že nemôžeme poznať všetky okolnosti, ktoré vplývajú na nastatie, či nenastatie sledovanej udalosti. Tak je to napr. pre udalosť „dnes poobede bude v Mlynskej doline pršať". Ak sme si nevzali dáždnik, tak máme šancu, povedzme 70 %-nú šancu, zmoknúť. Alebo napr. „dnešný zápas naše mužstvo na 90 % vyhrá". Všimnime si, že okrem vymedzenia akejsi udalosti, uvedené tvrdenia vyhodnocovali šancu, teda uvádzali mieru toho, že uvažovaná udalosť nastane.

Obe naznačené situácie – aj keď na pohľad celkom jasné – sú svojím spôsobom chúlostivé. Prečo? Nie je jednoduché povedať, odkiaľ v bežnom živote berieme hodnoty ako 70 %, resp. 90 %. Dohodnime sa, že predmetom našich úvah budú len také náhodné pokusy, pri ktorých

1. vieme špecifikovať podmienky, za ktorých sa pokus koná,
2. dokážeme určiť všetky (logicky možné) výsledky pokusu, ale nevieme s istotou predpovedať výsledok,
3. je možné (aspoň principiálne) pokus veľa ráz opakovať pri dodržaní podmienok pokusu,
4. výsledkom pokusu a náhodným udalostiam v ňom dokážeme priradiť pravdepodobnosť.

Teraz je zrejmé, že napr. body 1 a 3 v hore uvedených situáciách splnené nie sú. Naše úvahy budeme vzťahovať len na pokusy, v ktorých síce vystupuje „pani Náhoda", ale v ktorých sú skutočnosti 1 až 4 viac-menej splnené. Preto pôjde o situácie, v ktorých vieme určiť množinu Ω všetkých potenciálne možných výsledkov pokusu. Naučíme sa modelovať náhodné udalosti v pokuse ako podmnožiny množiny Ω. Systém všetkých náhodných udalostí v pokuse budeme označovať symbolom $\mathcal{S}$. Tým „orechovým" však bude pravdepodobnosť P, ktorej úlohou je vyhodnocovať udalosti v tom zmysle, že ak A je náhodná udalosť, tak P(A) je jej pravdepodobnosť, t. j. miera toho, že A nastane.

Trojicu $(\Omega, \mathcal{S}, \mathrm{P})$ nazývame *pravdepodobnostný priestor* a vo všeobecnej polohe ho predstavíme neskôr (v článku 1.4). Teraz začneme jednoduchými prípadmi priestoru, keď množina Ω všetkých (potenciálne možných) výsledkov pokusu je konečná.

Pri diskutovaní pojmu pravdepodobnosť sa často zabúda zdôrazniť, že jedna vec je vedieť určiť *správnu hodnotu* P(A) v konkrétnej situácii (a pre konkrétnu udalosť A) a druhá vec je pochopiť tie vlastnosti, ktoré má pravdepodobnosť ako *zobrazenie*. Vskutku, na P budeme nazerať ako na zobrazenie, ktoré udalostiam priraďuje číselnú hodnotu. V tejto kapitole preberieme tie vlastnosti, ktoré toto zobrazenie má v každom náhodnom pokuse.

1.1.1 Príklad. Predpokladajme, že v škatuli máme bielu, modrú a červenú guľku. Pokus spočíva v náhodnom ťahaní guľky, pričom za výsledok považujeme farbu vytiahnutej guľky. Hovorme o modeli takéhoto pokusu. Ako sme už povedali, symbol Ω bude predstavovať množinu všetkých výsledkov pokusu, teda teraz $\Omega = \{b, m, č\}$. Udalosti v pokuse sú nejaké výroky o tom, akými výsledkami pokus skončí, napr. udalosťou v pokuse je udalosť spočívajúca v tom, že vytiahnutá guľka nebude červená, t. j. že vytiahnutá bude biela alebo modrá. Označme túto udalosť symbolom A. Zrejme môžeme A chápať ako podmnožinu množiny Ω, teda ako množinu $\{b, m\}$, t. j. množinu pozostávajúcu práve z tých výsledkov pokusu, že keď nimi (práve nimi) pokus skončí, tak A nastane. Nie je ťažké vymenovať všetky možné udalosti pokusu a ako sme povedali, na označenie systému všetkých udalostí použijeme symbol $\mathcal{S}$. Teraz prvkami $\mathcal{S}$ sú množiny

$$\{b\},\ \{m\},\ \{č\},\ \{b, m\},\ \{b, č\},\ \{m, č\},\ \emptyset,\ \Omega\,.$$

Prvé tri udalosti nazývame *elementárne* udalosti. Môžeme ich stotožniť s výsledkami pokusu (rozdiel medzi výsledkami a nimi tvorenými elementárnymi udalosťami je len formálny). V našom (veľmi jednoduchom) pokuse máme iba tri udalosti, ktoré nie sú elementárne: $\{b, m\}$, $\{b, č\}$, $\{m, č\}$ a ešte dve špeciálne: *nemožnú* udalosť $\emptyset$ (ktorá nikdy nenastane) a udalosť *istú*, Ω (tá nastane vždy).

Teraz hovorme o pravdepodobnosti. V úvode kapitoly sme pravdepodobnosť uvádzali v percentách, čo sa v bežnom vyjadrovaní často robí. V matematike však pravdepodobnosť náhodnej udalosti je číslo z intervalu $\langle 0, 1\rangle$. Ak teda udalosť U znamená výhru (v nejakej hre s náhodou), tak skutočnosť, že $P(U) = 0.90$, je to isté, ako povedať, že pravdepodobnosť výhry je 90 %-ná. Stanovme teraz pravdepodobnosti náhodných udalostí nášho pokusu.

Začnime tým, že položíme $P(\emptyset) = 0$, $P(\Omega) = 1$ (lebo $\emptyset$ nikdy nenastáva a Ω nastáva vždy). Teraz určíme pravdepodobnosť elementárnych udalostí, teda jednotlivých výsledkov. Logika nám našepkáva, že v danej situácii jednotkovú pravdepodobnosť rozdelíme na tri tretiny, lebo každý z troch výsledkov je „rovnako možný" a všetky spolu vytvárajú istú udalosť. Takto dostávame $P(\{b\}) = P(\{m\}) = P(\{č\}) = \frac{1}{3}$.

Pretože udalosť $\{b, m\}$ nastáva práve keď je vytiahnutá biela alebo modrá guľka (sú to dve možnosti a nastáva vždy len jedna z nich), tak na základe intuície kladieme $P(\{b, m\}) = 2/3$. Analogicky určíme pravdepodobnosť aj ostatných dvojprvkových udalostí. Účinkovanie P na všetkých udalostiach môžeme zapísať tabuľkou, ktorá predstavuje pravdepodobnostný model pokusu:

$\mathcal{S}$	$\{b\}$	$\{m\}$	$\{č\}$	$\{b, m\}$	$\{b, č\}$	$\{m, č\}$	$\emptyset$	Ω
P	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	0	1

Poznamenávame, že sme položili $P(\{b, m\}) = P(\{b\}) + P(\{m\}) = 2/3$, odvolávajúc sa na intuíciu. Ide o prejav vlastnosti (najdôležitejšej vlastnosti) pravdepodobnosti, ktorej hovoríme *aditivita*. Je užitočné sledovať, ako budeme argumentovať pri obhajovaní toho, že *každá* pravdepodobnosť je aditívna.

1.1.2 Príklad. Nech pokus spočíva v náhodnom ťahaní guľky zo škatule, ktorá obsahuje 10 guliek: 5 bielych, 3 modré a 2 červené. Za výsledok opäť považujeme farbu vytiahnutej guľky. Čím sa takýto pokus líši od pokusu z 1.1.1? Zrejme Ω, $\mathcal{S}$ sú rovnaké ako v 1.1.1, ale P bude iná, veď vytiahnutie bielej má väčšiu pravdepodobnosť, ako vytiahnutie modrej, alebo červenej. Keďže vytiahnutie každej (z tých desiatich guliek) je rovnako možné, tak kladieme

$P(\{b\}) = 0.5$ (pretože 5 z desiatich realizujú udalosť $\{b\}$),

$P(\{m\}) = 0.3$ (pretože 3 z desiatich realizujú udalosť $\{m\}$),

$P(\{č\}) = 0.2$ (pretože 2 z desiatich realizujú udalosť $\{č\}$).

Takýto postup je zaiste v zhode s intuíciou, resp. so skúsenosťou. Môžeme povedať aj to, že postupujeme v zhode so známou Laplaceovou definíciou (mimochodom, s označením *definícia* sa dá v tomto prípade polemizovať), ktorú však vzťahujeme na pomocný priestor Ω_0, ktorý má 10 rovnako možných „prvotných výsledkov":

$$\Omega_0 = \{b1, b2, b3, b4, b5, m1, m2, m3, č1, č2\}.$$

Analogicky ako v 1.1.1 (odvolávajúc sa na aditivitu P), určíme pravdepodobnosti udalostí $A = \{b, m\}$, $B = \{b, č\}$, $C = \{m, č\}$. Teraz pravdepodobnosť P určuje tabuľka

$\mathcal{S}$	$\{b\}$	$\{m\}$	$\{č\}$	$\{b, m\}$	$\{b, č\}$	$\{m, č\}$	$\emptyset$	Ω
P	0.5	0.3	0.2	0.8	0.7	0.5	0	1

1.1.3 Príklad. Nech pokus spočíva v náhodnom ťahaní guľky zo škatule, ktorá obsahuje 100 guliek, avšak teraz vieme len to, že farby guličiek sú alebo biela, alebo modrá, alebo červená (teda nevieme, koľko ktorých). Výsledok pokusu je opäť farba vytiahnutej guľky. Čím sa takýto pokus líši od pokusu z 1.1.2? Zrejme Ω, $\mathcal{S}$ sú rovnaké, ale určiť P je niečo nové, lebo zloženie škatule nepoznáme. Ako teraz modelovať pravdepodobnosť spojenú s pokusom? Je nutné zamyslieť sa nad tým, čo je pravdepodobnosť. Na otázku, prečo veríme postupom v predchádzajúcich odstavcoch, odpovedáme takto:

Keby sme opakovane realizovali pokus z 1.1.2 (náhodne ťaháme guľku, poznačíme si jej farbu a vrátime ju späť), tak v dlhom rade opakovaní pokusu zistíme približnú zhodu medzi stanovenou pravdepodobnosťou P a relatívnymi početnosťami výskytu jednotlivých výsledkov, resp. udalostí. Ak totiž v 1.1.2 napr. $P(\{b\}) = 0.5$, tak v 1000 opakovaniach pokusu približne v polovici z nich, t. j. približne v 500 prípadoch, pozorujeme vytiahnutie bielej. Inými slovami, veríme, že frekvencia výsledku $\{b\}$ kolíše okolo 0.5.

Keďže ale teraz zloženie škatule nepoznáme, neostáva nič iné, len vykonať experiment. Opakujme náš pokus (napr. 60-krát) a predpokladajme, že sme získali takéto výsledky:

b b m b m b b m č b | *b b b č b č b m b b*
č b b m b b b m m b | *m b m b b č m č b b*
m č b b m b m m č m | *b m b b m b b m č b*

Overte, že absolútne početnosti jednotlivých výsledkov sú: 33, 18 a 9. Preto

relatívna početnosť výsledku b sa rovná 33/60 = 0.55,
relatívna početnosť výsledku m sa rovná 18/60 = 0.30,
relatívna početnosť výsledku $\check{c}$ sa rovná 9/60 = 0.15.

Vo všeobecnosti, po uskutočnení série n pokusov, relatívna početnosť výsledku b je podiel

$$\frac{k_n(b)}{n}$$

kde $k_n(b)$ je *absolútna* početnosť výsledku b v sérii n opakovaní pokusu. Všimnime si, že nech je n akékoľvek, čísla $\frac{k_n(b)}{n}$, $\frac{k_n(m)}{n}$, $\frac{k_n(\check{c})}{n}$ sú vždy z intervalu $\langle 0, 1 \rangle$ (čísla v čitateľoch sú zrejme menšie, nanajvýš sa rovnajú n). Tiež je zrejmé, že pre akékoľvek n platí

$$k_n(b) + k_n(m) + k_n(\check{c}) = n,$$

a preto

$$\frac{k_n(b)}{n} + \frac{k_n(m)}{n} + \frac{k_n(\check{c})}{n} = 1.$$

Ak $n \to \infty$, tak

$$\begin{array}{ccccccc} \downarrow & & \downarrow & & \downarrow & & \\ p_1 & + & p_2 & + & p_3 & = & 1. \end{array}$$

Ak teda n rastie nad všetky medze, tak experimenty nás presviedčajú o tom, že existujú hodnoty p_1, p_2, p_3, ktoré sú (niečo ako) limity relatívnych početností a môžeme ich považovať za pravdepodobnosti výsledkov pokusu. Zrejme platí $0 \le p_1, p_2, p_3 \le 1$ a pre súčet máme: $p_1 + p_2 + p_3 = 1$.

Ďalej si všimnime, že pre akékoľvek n, pre absolútnu početnosť udalosti $\{b, m\}$ platí

$$k_n(\{b,m\}) = k_n(\{b\}) + k_n(\{m\}),$$

a preto rovnosť

$$\frac{k_n(\{b,m\})}{n} = \frac{k_n(\{b\})}{n} + \frac{k_n(\{m\})}{n}$$

platí pre každé n. To je dôvod, prečo pre P žiadame platnosť vzťahu

$$\mathrm{P}(\{b,m\}) = \mathrm{P}(\{b\}) + \mathrm{P}(\{m\}).$$

Hovoríme, že P je *aditívna*. Aditivita, vo všeobecnosti, znamená:

$$\boxed{\text{Ak udalosti A, B sú disjunktné, t. j. ak } A \cap B = \varnothing, \text{ potom } \mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B).}$$

(V našej situácii $A = \{b\}$, $B = \{m\}$.)

Je dôležité uvedomiť si, že

- frekvenčná interpretácia pravdepodobnosti má v inžinierskych aplikáciách zásadný význam (spomedzi ďalších interpretácií je najdôležitejšia),
- frekvencie udalostí, t. j. relatívne početnosti udalostí, majú vlastnosť aditivity, a preto je celkom prirodzené, že vlastnosť aditivity pripisujeme aj samotnej pravdepodobnosti P.

Zakončime príklad 1.1.3. Uvedomili sme si, že v ňom nie je možné modelovať pravdepodobnosť špekulatívnou metódou. V skutočnosti nepoznáme pravdepodobnosti výsledkov b, m, $\check{c}$ môžeme len odhadnúť. Ak relatívnymi početnosťami odhadneme p_1, p_2, p_3 (pritom tak, aby platilo: $p_1 + p_2 + p_3 = 1$), potom (vďaka aditivite) vieme stanoviť pravdepodobnosť akejkoľvek udalosti. Pravdepodobnostnú štruktúru priestoru popisuje tabuľka:

S	$\{b\}$	$\{m\}$	$\{\check{c}\}$	$\{b, m\}$	$\{b, \check{c}\}$	$\{m, \check{c}\}$	$\varnothing$	Ω
P	p_1	p_2	p_3	$p_1 + p_2$	$p_1 + p_3$	$p_2 + p_3$	0	1

Celkom konkrétne, ak napr. po 1000-násobnom opakovaní pokusu dostaneme $k_{1000}(\{b\}) = 507$, $k_{1000}(\{m\}) = 297$ a $k_{1000}(\{\check{c}\}) = 196$, tak za model pokusu berieme model daný tabuľkou:

S	$\{b\}$	$\{m\}$	$\{\check{c}\}$	$\{b, m\}$	$\{b, \check{c}\}$	$\{m, \check{c}\}$	$\varnothing$	Ω
P	0.507	0.297	0.196	0.804	0.703	0.493	0	1

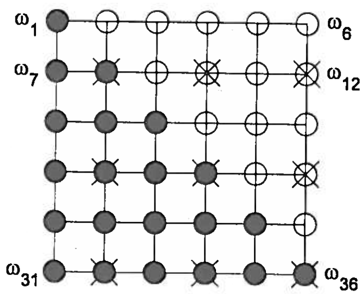
Ak vezmeme do úvahy fakt, že škatuľa obsahovala 100 guliek, tak naše tipy na zloženie škatule by boli: 50 – 30 – 20, resp. 51 – 29 – 20 a pod. Avšak 100 %-ný tip zrejme neexistuje.

1.1.4 Príklad. Nech pokus spočíva v hode dvoma riadnymi hracími kockami, modrou a červenou. Za výsledok pokusu považujme usporiadanú dvojicu čísel (i, j), ktoré padli na kockách (i bodov na modrej, j na červenej). Uvažujme o udalostiach A, B:

A – spočíva v tom, že na modrej padne aspoň toľko bodov ako na červenej,
B – spočíva v tom, že na každej kocke padne párne číslo.

Modelujme ich ako podmnožiny množiny všetkých výsledkov. Všimnime si, že systém náhodných udalostí má istú štruktúru, čo znamená, že s dvojicou udalostí A, B sa prirodzene objavujú ďalšie udalosti: $A \cap B$, $A \cup B$, $A \setminus B$, resp. $B \setminus A$. Nakoniec, definujme pravdepodobnosť v tomto pokuse.

Riešenie. Zrejme množina Ω (ktorá modeluje všetky možné výsledky pokusu) má 36 prvkov. Prvky ω množiny Ω sú usporiadané dvojice (i, j) a môžeme ich znázorniť ako uzlové body mriežky (obr. 1-1) a očíslovať zľava doprava a zhora nadol: $\omega_1 = (1, 1)$, $\omega_2 = (1, 2)$, ..., $\omega_6 = (1, 6)$, $\omega_7 = (2, 1)$, $\omega_8 = (2, 2)$, ..., $\omega_{12} = (2, 6)$, ..., $\omega_{31} = (6, 1)$, $\omega_{32} = (6, 2)$, ..., $\omega_{36} = (6, 6)$.



Obr. 1-1. Množina Ω z príkladu 1.1.4

Udalosť A modelujeme množinou
$A = \{(i, j): i \ge j,\ i, j = 1, 2, \dots, 6\}$
(body A sú znázornené ako plné uzly).

Udalosť B modelujeme množinou
$B = \{(i, j): i, j \in \{2, 4, 6\}\}$
(prvky B sú označené krížikom).

S udalosťami A, B sa vynárajú nové, ktoré definujeme pomocou A, B, napr. $A \cap B$, $A \cup B$, $A \setminus B$, resp. $B \setminus A$.

$A \cap B$ ako množina je dobre známa, $A \cap B = \{\omega_8, \omega_{20}, \omega_{22}, \omega_{32}, \omega_{34}, \omega_{36}\}$. Ako udalosť ju môžeme opísať tak, že nastáva práve vtedy, keď v pokuse nastane aj udalosť A, aj udalosť B.

$A \cup B$ ako množina je opäť každému známa, a pretože má až 24 prvkov, nevypisujeme ich. Ako udalosť ju môžeme definovať tak, že nastáva práve vtedy, keď v pokuse nastane aspoň jedna z udalostí A, B.

$A \setminus B$ (množinový rozdiel) je v jazyku udalostí taká udalosť, ktorá nastáva práve vtedy, keď v pokuse nastane A a súčasne udalosť B nenastane.

Operácia množinového komplementu umožňuje hovoriť o opačnej udalosti, napr. ak A je hore uvažovaná udalosť $\{(i,j) : i \geq j,\ i,j = 1,2,\dots,6\}$, (na modrej aspoň toľko bodov, ako na červenej), tak opačná udalosť spočíva v tom, že na červenej kocke padne viac bodov, teda $A' = \{(i,j) : i < j,\ i,j = 1, 2, \dots, 6\}$. Použitím operácie komplementu môžeme vyjadriť $A \setminus B$, pretože zrejme platí $A \setminus B = A \cap B'$. Samozrejme, $B \setminus A = B \cap A'$.

Po stotožnení udalostí s podmnožinami množiny Ω, je systém všetkých udalostí v pokuse vlastne systémom všetkých podmnožín množiny Ω. Poznamenávame, že v situáciách, keď Ω je konečná, nič nám nebráni za systém *všetkých* udalostí v pokuse zobrať vždy systém *všetkých* podmnožín množiny Ω.

Hovorme teraz o pravdepodobnosti v našom pokuse. Pretože kocky sú riadne, t. j. sú to presné kocky (v geometrickom slova zmysle), vyrobené z homogénneho materiálu, považujeme všetkých 36 možných výsledkov za rovnako pravdepodobné výsledky a každému z nich prisudzujeme pravdepodobnosť 1/36. Ak napr. $D_2 = \{\omega_1, \omega_2\}$, tak vďaka aditivite $P(D_2) = P(\{\omega_1, \omega_2\}) = P(\{\omega_1\}) + P(\{\omega_2\}) = 2/36$, (pozri 1.1.3). Ak napr. $D_3 = \{\omega_1, \omega_2, \omega_3\}$, tak

$$P(D_3) = P(D_2 \cup \{\omega_3\}) = P(D_2) + P(\{\omega_3\}) = 2/36 + 1/36 = 3/36.$$

Opakovaním tohoto postupu dostaneme napr. pre $D_6 = \{\omega_1, \omega_2, \dots, \omega_6\}$, $P(D_6) = 6/36$. Takto z aditivity máme $P(A) = 21/36$, pretože $|A| = 21$ (počet prvkov A sa rovná 21) a analogicky, $P(B) = 9/36$, pretože $|B| = 9$.

1.1.5 Hore uvedené postupy, resp. argumenty z príkladu 1.1.4 môžeme uplatniť v každom pokuse, v ktorom je Ω konečná a jej body (t. j. výsledky pokusu) sú rovnako pravdepodobné. Ukážeme, že odvolávanie sa na aditivitu dáva to isté, ako odvolávanie sa na známu Laplaceovu definíciu pravdepodobnosti.

Nech $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ a nech $P(\omega_i) = a$, pre všetky $i = 1, 2, \dots, n$. Opakovane uplatňujme aditivitu:

$$1 = P(\Omega) = P(\{\omega_1\} \cup \{\omega_2, \omega_3, \dots, \omega_n\}) = P(\{\omega_1\}) + P(\{\omega_2, \omega_3, \dots, \omega_n\}) =$$

$$= a + P(\{\omega_2\} \cup \{\omega_3, \omega_4, \dots, \omega_n\}) = a + a + P(\{\omega_3, \omega_4, \dots, \omega_n\}) = \dots = a + a + \dots + P(\{\omega_n\}) = na,$$

odkiaľ zrejme $a = 1/n$. Ak k bodov množiny Ω vytvára udalosť A, tak argumentovaním aditivitou, po $k - 1$ krokoch dostaneme $P(A) = \frac{k}{n}$.

To isté hovorí klasická definícia pravdepodobnosti (Laplaceova definícia): Nech pokus má konečne veľa výsledkov, t. j. $|\Omega| = n$, pričom všetky výsledky sú rovnako pravdepodobné. Ak k výsledkov je *priaznivých* pre nastatie A, t. j. $|A| = k$, (A pozostáva z k prvkov), tak pre P(A) platí

$$P(A) = \frac{k}{n}$$

Striktne vzaté, Laplaceova definícia nie je definícia pravdepodobnosti (ako takej), ale je to návod, ako nájsť pravdepodobnosť konkrétnej udalosti A v špeciálnej situácii – ak je výsledkov pokusu konečne veľa a ak sú všetky rovnako pravdepodobné.

Adekvátne použitie tohoto návodu teda vyžaduje istú intuíciu – je treba správne postrehnúť, kedy výsledky pokusu sú, resp. kedy nie sú rovnako pravdepodobné.

Úlohy

1.1.1 Nech A, B, C sú náhodné udalosti. Pomocou A, B, C vyjadrite udalosť spočívajúcu v tom, že

a) nastane len udalosť A (teda spomedzi A, B, C nastane len A),
b) nastane A a B, a pritom C nenastane,
c) nastanú všetky tri udalosti (teda A, B, C nastanú súčasne),
d) nastane práve jedna z nich,
e) nastane aspoň jedna z nich,
f) nastanú práve dve z nich,
g) nenastane žiadna z nich.

1.1.2 Náhodný pokus spočíva v hode tromi kockami: modrou, červenou a žltou. Zvoľme náhodné udalosti takto:

M_i na modrej padne i bodov ($i = 1, 2, \dots, 6$)
C_j na červenej padne j bodov ($j = 1, 2, \dots, 6$)
Z_k na žltej padne k bodov ($k = 1, 2, \dots, 6$).

Pomocou M_i, C_j, Z_k vyjadrite udalosti:

a) na modrej a aj na červenej padne aspoň 5 bodov,
b) na každej kocke padne párne číslo,
c) na všetkých padne to isté číslo,
d) aspoň na jednej padne 6,
e) aspoň na dvoch padne 6.

1.1.3 Náhodný pokus spočíva v hode tromi kockami: modrou, červenou a žltou. Za výsledok pokusu považujme usporiadanú trojicu (i, j, k) bodov, ktoré padli na modrej, červenej, resp. žltej. Množina Ω všetkých možných výsledkov má 216 prvkov a zrejme všetky výsledky sú rovnako pravdepodobné. Označme náhodné udalosti ako v predchádzajúcej úlohe:

M_i na modrej padne i bodov,
C_j na červenej padne j bodov,
Z_k na žltej padne k bodov.

Vyjadrite udalosti M_i, C_j, Z_k ako podmnožiny množiny Ω. Ďalej zistite, koľko prvkov majú udalosti A, B, ..., E a určte ich pravdepodobnosti.

a) $A = (M_5 \cup M_6) \cap (C_5 \cup C_6)$.

b) $B = (M_2 \cup M_4 \cup M_6) \cap (C_2 \cup C_4 \cup C_6) \cap (Z_2 \cup Z_4 \cup Z_6)$.

c) $C = (M_1 \cap C_1 \cap Z_1) \cup (M_2 \cap C_2 \cap Z_2) \cup (M_3 \cap C_3 \cap Z_3) \cup \dots \cup (M_6 \cap C_6 \cap Z_6)$.

d) $D = M_6 \cup C_6 \cup Z_6$.

e) $E = (M_6 \cap C_6) \cup (M_6 \cap Z_6) \cup (C_6 \cap Z_6)$.

1.1.4 Náhodný pokus spočíva v hode tromi označenými mincami. Označme A_i – na i-tej minci padol znak, $i = 1, 2, 3$. Pomocou A_1, A_2, A_3 vyjadrite náhodné udalosti

B – znak padne len na prvej a tretej minci, C – padnú práve dva znaky,

D – padnú najviac dva znaky, E_i – znak padne len na i-tej minci.

1.1.5 Náhodný pokus spočíva v hode tromi označenými mincami. Za výsledok pokusu považujeme usporiadanú trojicu symbolov Z (znak), C (číslo). Zrejme množina Ω všetkých možných výsledkov má 8 bodov. Vyjadrite udalosti z predchádzajúcej úlohy ako podmnožiny množiny Ω a určite ich pravdepodobnosti.

1.1.6 Systém pozostáva z dvoch blokov typu I a troch blokov typu II. Náhoda ovplyvňuje fungovanie, resp. nefungovanie jednotlivých blokov. Označme udalosti takto:

A_i i-tý blok typu I funguje, B_j j-tý blok typu II funguje.

Nasledujúce udalosti zapíšte pomocou A_i, B_j

a) udalosť C spočívajúcu v tom, že funguje len druhý blok typu I a len tretí blok typu II,

b) udalosť D núdzového režimu, ktorý nastáva, ak funguje práve jeden blok typu I a súčasne práve jeden blok typu II,

c) udalosť E spoľahlivého režimu, ktorý vyžaduje fungovanie aspoň jedného bloku typu I a súčasne fungovanie aspoň dvoch blokov typu II.

1.1.7 Systém pozostáva z dvoch blokov typu I a troch blokov typu II. Náhoda ovplyvňuje fungovanie jednotlivých blokov. Možné stavy systému považujeme za možné výsledky experimentu s náhodou a jednotlivé stavy (teda výsledky) môžeme zachytiť usporiadanými 5-ticami núl a jednotiek. Napr. stav, keď funguje práve druhý blok typu I a prvé dva bloky typu II označíme 5-ticou (0, 1, 1, 1, 0).

Nech A_i, B_j sú náhodné udalosti z predchádzajúcej úlohy. Zistite, z koľkých stavov pozostávajú nasledujúce udalosti

$$C = A_1' \cap A_2 \cap B_1' \cap B_2' \cap B_3,$$

$$D = (A_1 \cap A_2' \cap B_1 \cap B_2' \cap B_3') \cup (A_1 \cap A_2' \cap B_1' \cap B_2 \cap B_3') \cup (A_1 \cap A_2' \cap B_1' \cap B_2' \cap B_3) \cup$$
$$\cup (A_1' \cap A_2 \cap B_1 \cap B_2' \cap B_3') \cup (A_1' \cap A_2 \cap B_1' \cap B_2 \cap B_3') \cup (A_1' \cap A_2 \cap B_1' \cap B_2' \cap B_3),$$

$$E = (A_1 \cup A_2) \cap [(B_1 \cap B_2) \cup (B_1 \cap B_3) \cup (B_2 \cap B_3)].$$

Určite P(C), P(D), P(E), ak všetky stavy sú rovnako pravdepodobné.

1.2 Uplatnenie kombinatoriky pri určovaní P(A)

1.2.1 V tomto článku pôjde o také náhodné pokusy, aké sme diskutovali v 1.1.4 a 1.1.5. Množina všetkých výsledkov pokusu bude konečná, $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$, a pritom pravdepodobnosť každého výsledku bude rovnaká, t. j. $P(\omega_i) = 1/n$, pre $i = 1, 2, \dots, n$. Vďaka aditivite (ako sme ukázali v 1.1.5), ak udalosť A má k prvkov, tak pre P(A) máme

$$P(A) = \frac{k}{n}, \quad \text{t.j. } P(A) = \frac{|A|}{|\Omega|}.$$

Tentoraz však číslo n bude veľké a vtedy na zistenie počtu prvkov množín (či už Ω, alebo A) je možné, resp. nutné, využiť poznatky kombinatoriky. Predtým, než sa pustíme do zopakovania základných myšlienok kombinatoriky, všimnime si, že pre konečné množiny A, B a pre počty ich prvkov $|A|, |B|$ platí:

$$|A| + |B| = |A \cup B| + |A \cap B|,$$

a preto tiež

$$\frac{|A|}{|\Omega|} + \frac{|B|}{|\Omega|} = \frac{|A \cup B|}{|\Omega|} + \frac{|A \cap B|}{|\Omega|},$$

to znamená, že

$$P(A) + P(B) = P(A \cup B) + P(A \cap B).$$

Ako uvidíme neskôr, platnosť tohoto vzťahu je všeobecná, aj keď sme ho zatiaľ zdôvodnili iba pre prípad *homogénneho* priestoru (t. j. priestoru, v ktorom výsledkov pokusu je konečne veľa a všetky majú rovnakú pravdepodobnosť). Najčastejšie ho využívame v tvare

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Ak platí $A \cap B = \emptyset$, tak $P(A \cap B) = 0$ a máme $P(A \cup B) = P(A) + P(B)$, čo je známa aditivita P.

1.2.2 Kombinatorické pojmy a výsledky preberieme v kontexte dvoch najčastejších situácií – buď pôjde o *náhodné výbery*, alebo o *náhodné rozmiestnenia*. Najprv hovorme o náhodných výberoch. Predstavme si, že v škatuli máme n rozlíšiteľných (napr. očíslovaných) loptičiek a z nich náhodne vyberieme k (číslu k hovoríme *rozsah* výberu). Réžia pokusu môže byť rôzna – náhodne vyberieme prvú a vrátime ju späť pred ťahaním druhej (resp. ďalšej), alebo, naopak, vytiahnuté loptičky sa naspäť dávať nebudú. Hovoríme, že buď ide o *výber s vrátením*, alebo *výber bez vrátenia*. Na druhej strane, za výsledok výberu môžeme považovať buď *usporiadanú* k-ticu loptičiek (teda vlastne čísel), alebo *neusporiadanú* k-ticu. Takto pôjde buď o usporiadané, alebo neusporiadané výbery. Venujme sa postupne jednotlivým prípadom.

1.2.3 (*Usporiadaný výber s vrátením*) V škatuli máme n rozlíšiteľných loptičiek (povedzme, že sú očíslované číslami $1, 2, \dots, n$) a ide o výber s vrátením. To znamená, že si poznamenáme číslo vybratej a vrátime ju späť pred ťahaním ďalšej. Za výsledok celého výberu rozsahu k považujeme usporiadanú k-ticu čísel z množiny $\{1, 2, \dots, n\}$. Pretože ide o výber s vrátením, rozsah výberu k môže byť ľubovoľné prirodzené číslo – to znamená, že môže byť $k < n$, $k = n$, ale aj $k > n$.

Koľko je možných výberov? Výsledkom výberu je usporiadaná k-tica čísel z $\{1, 2, \dots, n\}$

$$(x_1, x_2, \dots, x_k), \quad x_i \in \{1, 2, \dots, n\}, \quad i = 1, 2, \dots, k.$$

Na prvom mieste v k-tici môže byť akékoľvek číslo z $\{1, 2, \dots, n\}$, na druhom mieste opäť, pretože prvá vytiahnutá loptička sa vrátila späť pred ťahaním druhej. Ak $k = 2$, máme $nn = n^2$ dvojíc. Ak $k = 3$, tak každá dvojica z n^2 dvojíc zrodí n rôznych trojíc, a preto počet usporiadaných trojíc sa rovná $n^2 n = n^3$. Pokračujúc v úvahe, počet všetkých možných usporiadaných k-tíc sa rovná

$$n^k$$

V kombinatorickej terminológii hovoríme: počet *variácií s opakovaním k-tej triedy z n prvkov* sa rovná n^k.

Nech $n = 4$, $k = 5$. Možnými výsledkami výberu sú 5-tice: (3, 3, 3, 3, 3), alebo (4, 2, 4, 2, 1) a všetkých možných výberov je $4^5 = 2^{10} = 1024$. Aká je pravdepodobnosť udalosti A, ktorá spočíva v tom, že vo výbere (t. j. vo vybratej 5-tici) budú iba párne čísla?

Počet usporiadaných 5-tíc párnych čísel, teda čísel z množiny $\{2, 4\}$ sa rovná 2^5. Preto $|A| = 32$. Takto máme $P(A) = 32/1024$.

1.2.4 (*Usporiadaný výber bez vrátenia*) Znovu ide o usporiadaný výber, tentoraz však vybraté loptičky do škatule nevraciame. Výsledkom je opäť usporiadaná k-tica, avšak teraz sa v nej žiadne číslo neopakuje, keďže raz vytiahnutá loptička už vytiahnutá byť nemôže. V tomto prípade, samozrejme, musí nutne platiť $k \le n$. Koľko je možných výberov?

Ide znovu o usporiadané k-tice. Na prvom mieste môže figurovať ľubovoľné z n čísel, avšak na druhom mieste ľubovoľné spomedzi $n - 1$ čísel, pretože prvá vytiahnutá sa späť nevrátila. Počet všetkých možných k-tíc sa v tomto prípade rovná súčinu k činiteľov, ktorý označujeme $V(n, k)$

$$V(n, k) = n(n-1)(n-2)\dots(n-(k-1)) = n(n-1)(n-2)\dots(n-k+1)$$

V kombinatorike hovoríme: počet *variácií bez opakovania k-tej triedy z n prvkov* sa rovná súčinu $n(n-1)(n-2)\dots(n-k+1)$. Ak napr. $n = 4$, $k = 3$, tak $V(n, k) = V(4, 3) = 4 \cdot 3 \cdot 2 = 24$.

1.2.5 Príklad. V škatuli máme 10 rozlíšiteľných výrobkov. Realizujme usporiadaný náhodný výber rozsahu 5 s vrátením. Aká je pravdepodobnosť udalosti A spočívajúcej v tom, že vo výbere sa žiaden výrobok neobjaví viac ako raz?

Riešenie. Zrejme počet všetkých možných výberov sa rovná 10^5 (podľa 1.2.3). Počet výberov priaznivých pre nastatie A sa rovná $V(10, 5) = 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 = 30\,240$, a preto

$$P(A) = 30\,240/100\,000 = 0.3024.$$

1.2.6 (*Náhodné permutácie*) Ostávame v kontexte odstavca 1.2.4 a nech platí $k = n$. Výsledok výberu môžeme interpretovať ako náhodnú permutáciu n-tice $(1, 2, 3, \dots, n)$. Ak sú loptičky očíslované číslami 1 až n, tak po uskutočnení náhodného výberu rozsahu n bez vrátenia, máme k dispozícii n-ticu

$$(x_1, x_2, \dots, x_n), \quad x_i \in \{1, 2, \dots, n\}, \quad i = 1, \dots, n, \quad x_i \ne x_j, \text{ pre } i \ne j, \quad i, j = 1, 2, \dots, n,$$

ktorá predstavuje náhodné preusporiadanie základného poradia, t. j. n-tice $(1, 2, \dots, n)$.

Počet možných výberov (t. j. možných permutácií) sa rovná súčinu $n(n-1)(n-2)\dots 3 \cdot 2 \cdot 1$, ktorý označujeme známym symbolom $n!$.

1.2.7 Príklad. Uvažujme o náhodných permutáciách prvkov $1, 2, \dots, 9$. Nech udalosť A vytvárajú práve také permutácie, v ktorých prvok 5 je na svojom mieste (t. j. prvok 5 je piatou zložkou 9-tice). Aká je pravdepodobnosť udalosti A, ak všetky permutácie považujeme za rovnako pravdepodobné? Mimochodom, keď vznikajú tak, ako je opísané v odseku 1.2.4, tak o tom zrejme nepochybujeme.

Riešenie. Všetkých permutácií je 9!. Koľko prvkov má A? Prvok 5 musí byť na svojom mieste a ostatných 8 prvkov môže ležať na ľubovoľných miestach, preto $|A| = 8!$. Napr. permutácia (1, 2, 3, 6, 5, 4, 7, 8, 9) tiež realizuje udalosť A. Preto $P(A) = 8!/9! = 1/9$.

Stojí za zmienku, že ak namiesto A, uvažujeme udalosť B, ktorá spočíva v tom, že iba prvok 5 je na svojom mieste, tak určenie P(B) je iná úloha. Napríklad, hore uvedená permutácia, ktorá realizovala A, samozrejme, nerealizuje B. Udalosť B realizuje napríklad permutácia (2, 1, 4, 7, 5, 8, 9, 3, 6), pretože jediným číslom na svojom mieste je 5. Nájsť P(B) je trochu ťažšie, ako určiť P(A). Urobíme to neskôr.

1.2.8 (*Neusporiadaný výber bez vrátenia*) Tentoraz nebudeme rozlišovať medzi výbermi, keď sa budú líšiť len poradím vytiahnutých prvkov (čísel). Ak napr. $n = 4$, $k = 3$, tak v odstavci 1.2.4 sme brali napr. trojice (1, 3, 4) (1, 4, 3) (3, 1, 4) (3, 4, 1) (4, 1, 3) (4, 3, 1) ako 6 rôznych výberov, ale teraz ich stotožňujeme a chápeme ako jeden: $\{1, 3, 4\}$. Pretože v 1.2.4 sme uvažovali k-tice pozostávajúce z *rôznych* prvkov, tak vždy $k!$ prípadov z 1.2.4 teraz chápeme ako jeden. Preto počet všetkých neusporiadaných výberov bez vrátenia sa rovná

$$\frac{V(n, k)}{k!}$$

V kombinatorike hovoríme: počet *kombinácií k-tej triedy z n prvkov* (*bez opakovania*) sa rovná $V(n, k)/n!$ a označujeme $C(n, k)$, resp. ako kombinačné číslo „n nad k"

$$\binom{n}{k} = \frac{n(n-1)(n-2)\dots(n-k+1)}{k!} = \frac{n!}{(n-k)!\,k!} = C(n, k)$$

Konkrétna kombinácia k-tej triedy nie je nič iné, ako konkrétna k prvková podmnožina n prvkovej množiny. Hore uvedený výsledok (o počte kombinácií, teda o počte možných neusporiadaných výberov bez vrátenia) môžeme interpretovať tak, že počet k prvkových podmnožín danej n prvkovej množiny sa rovná $C(n, k)$.

Vo svetle výberov sú niektoré fakty okolo kombinačných čísel celkom zjavné, napr. skutočnosť, že

$$C(n, k) = C(n, n-k), \text{ t. j. že platí: } \binom{n}{k} = \binom{n}{n-k}.$$

Zrejme vybrať k prvkov z n prvkov je to isté, ako určiť tých $n - k$ prvkov, ktoré do tej vytváranej k prvkovej podmnožiny nevezmeme.

Ďalej, ak v známej binomickej vete položíme $a = 1$, $b = 1$, máme

$$(a+b)^n = (1+1)^n = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \dots + \binom{n}{n-1} + \binom{n}{n} = 1 + \binom{n}{1} + \binom{n}{2} + \dots + \binom{n}{n-1} + 1$$

Pravú stranu môžeme chápať ako súčet počtov všetkých podmnožín n prvkovej množiny: prvá a posledná jednotka sú pre prázdnu podmnožinu a nevlastnú podmnožinu. Kombinačné čísla predstavujú počty podmnožín: postupne jednoprvkových, dvojprvkových atď. Takto dostávame zdôvodnenie známeho faktu:

Počet všetkých podmnožín množiny, ktorá má n prvkov, sa rovná 2^n.

1.2.9 Príklad. V škatuli máme 30 výrobkov a medzi nimi je 5 nepodarkov. Náhodne vyberme naraz 4. Aká je pravdepodobnosť toho, že medzi vybratými bude práve jeden nepodarok?

Riešenie. Vybrať naraz štyri je to isté, ako realizovať neusporiadaný výber rozsahu 4 bez vrátenia. Počet možných výberov sa rovná $C(30, 4) = 27\,405$ a všetky sú rovnako pravdepodobné. Nech A je udalosť, že medzi vybratými je jeden nepodarok. Koľko prvkov má množina A?

Jeden nepodarok z piatich nepodarkov možno vybrať piatimi spôsobmi. Počet možností ako vybrať 3 dobré výrobky z 25 dobrých sa rovná $C(25, 3) = 2300$. Každý z tých piatich spôsobov vytvorí s každou z tých 2300 možností priaznivú štvoricu pre nastatie A. Preto počet prvkov A sa rovná $5 \cdot C(25, 3) = 5 \cdot 2300 = 11\,500$. Nakoniec

$$P(A) = \frac{\binom{5}{1}\binom{25}{3}}{\binom{30}{4}} = \frac{11500}{27405} = 0.4196.$$

1.2.10 Príklad. V škatuli máme loptičky očíslované číslami 1 až 9. Realizujme náhodný výber rozsahu 3 bez vrátenia a za výsledok považujme usporiadanú trojicu čísel. Aká je pravdepodobnosť toho, že výsledok výberu vytvorí rastúcu postupnosť?

Riešenie. Počet všetkých možných usporiadaných výberov bez vrátenia sa rovná súčinu $9 \cdot 8 \cdot 7 = 504$ (podľa 1.2.4). Nech A znamená udalosť, že výber skončí rastúcou postupnosťou troch čísel. Treba nájsť $|A|$.

Pretože ide o trojice *rôznych* prvkov, vždy šiestim prípadom odpovedá jeden, priaznivý pre udalosť A. Napr. trojiciam (1, 3, 4) (1, 4, 3) (3, 1, 4) (3, 4, 1) (4, 1, 3) (4, 3, 1) odpovedá jedna rastúca trojica (1, 3, 4). Keďže počet všetkých usporiadaných trojíc sa rovná $V(9, 3)$, počet rastúcich trojíc sa rovná $V(9, 3)/3! = C(9, 3)$. To znamená, že $|A| = C(9, 3) = 84$. Pre P(A) máme

$$P(A) = \frac{84}{504} = \frac{1}{6}$$

1.2.11 (*Náhodné rozmiestnenia – mechanizmus*) Uvažujme o n priečinkoch, očíslovaných od 1 po n. Majme k obálok, ktoré budeme náhodne rozmiestňovať do priečinkov. Čo znamená náhodne? Predstavme si, že bokom umiestnime škatuľu, ktorá obsahuje n lístkov s číslami 1 až n. Náhodne ťaháme *prvý* lístok zo škatule a číslo na lístku bude znamenať číslo priečinku, do ktorého dáme *prvú* obálku. Druhá obálka bude putovať do priečinka s takým číslom, aké bolo uvedené na druhom vytiahnutom lístku atď. Zrejme na umiestnenie k obálok bude potrebný náhodný výber rozsahu k.

Budeme diskutovať prípad, keď obálky sú rozlíšiteľné (označené) a prípad, keď nie sú. Ďalšou okolnosťou bude, či priečinok môže obsahovať viac ako jednu obálku, alebo nie. Rozoberme jednotlivé prípady.

1.2.12 (*Rozmiestnenie k rozlíšiteľných obálok do n priečinkov bez zákazu*) Skratkou „bez zákazu" budeme označovať prípad, keď priečinok môže obsahovať viac obálok, hoci aj všetky. Koľko je v takomto prípade rôznych rozmiestnení? Najprv si ujasnime, kedy ide o dve rôzne rozmiestnenia. Prirodzene vtedy, ak existuje aspoň jeden priečinok, ktorého obsah je iný v prvom a iný pri druhom rozmiestnení. To však znamená, že existuje obálka (obálka napr. s číslom l), ktorá pri prvom rozmiestnení putovala do priečinku i a v druhom rozmiestnení do priečinku j (samozrejme, $i \neq j$). Prvá obálka môže byť umiestnená do ľubovoľného priečinku – to je n možností. Druhá obálka opäť do ľubovoľného priečinku (ide teraz o prípad bez zákazu), takže počet umiestnení dvoch *rozlíšiteľných* obálok sa rovná n^2. Pokračujúc v úvahe, dostávame nakoniec, že počet rozmiestnení bez zákazu k rozlíšiteľných obálok do n priečinkov sa rovná n^k.

Ako by sme realizovali náhodné rozmiestnenie tohoto typu? To sme opísali v odstavci 1.2.11 – rozmiestňovanie môže byť realizované prostredníctvom náhodného výberu lístka zo škatule. Len treba upresniť, že teraz by išlo o usporiadaný výber s vrátením. Je dobré mať na mysli jedno-jednoznačnú korešpodenciu (t.j. bijekciu) medzi usporiadanými výbermi s vrátením a rozmiestneniami rozlíšiteľných obálok bez zákazu. Tá korešpodencia (bijekcia) je definovaná postupom, mechanizmom rozmiestňovania, opísanom v 1.2.11.

1.2.13 (*Rozmiestnenie k rozlíšiteľných obálok do n priečinkov so zákazom*) Teraz priečinok nesmie obsahovať viac ako jednu obálku (teda jednu, alebo žiadnu). Musí platiť $k \leq n$, lebo inak sa rozmiestnenie nedá realizovať. Koľko je teraz možných rozmiestnení? Prvá obálka môže putovať do ľubovoľného priečinku, druhá už len do ľubovoľného z $n - 1$ priečinkov, tretia už len do ľubovoľného z $n - 2$ priečinkov atď. Zrejme počet všetkých rozmiestnení sa rovná $V(n, k)$. Realizovať náhodné rozmiestnenie tohoto typu môžeme pomocou mechanizmu z 1.2.11, ale teraz by išlo o výbery *bez vrátenia* (aby išlo o rozmiestňovanie so zákazom).

1.2.14 Príklad. (*Narodeninový paradox*) Predstavme si, že v triede je 30 žiakov a nie sú v nej dvojičky. Aká je pravdepodobnosť p toho, že sú medzi žiakmi takí dvaja, ktorí majú narodeniny v ten istý deň?

Riešenie. Predstavme si 365 priečinkov (čo priečinok, to jeden deň v roku). Pani Náhoda pred rokmi náhodne rozmiestnila 30 rozlíšiteľných loptičiek do tých 365 priečinkov. Otázka je, aká je pravdepodobnosť toho, že existuje priečinok, ktorý obsahuje viac ako jednu loptičku. Jednoduchšie je však najprv nájsť pravdepodobnosť toho, že taký priečinok neexistuje.

Uvážme, že množina všetkých možných rozmiestnení loptičiek do priečinkov má 365^{30} prvkov. To preto, lebo prvá loptička mohla skončiť v ľubovoľnom priečinku, druhá tiež v hociktorom z 365 priečinkov atď. a to je $365 \cdot 365 \cdot 365 \cdot \ldots \cdot 365 = 365^{30}$. Koľko z týchto rozmiestnení je takých, že žiaden priečinok neobsahuje dve a viac loptičiek? Takých je

$$365 \cdot 364 \cdot 363 \cdot \ldots \cdot (365 - 29) = 365 \cdot 364 \cdot 363 \cdot \ldots \cdot 336 = V(365, 30)$$

To preto, že kým prvá loptička mohla skončiť v ľubovoľnom priečinku, druhá mala len 364 možností, tretia len 363 atď. Preto pravdepodobnosť q toho, že žiaden priečinok neobsahuje viac ako jednu loptičku sa rovná

$$q = \frac{365\,364\,363\,...\,336}{365^{30}}$$

Opačnou udalosťou je, že existuje priečinok, ktorý obsahuje viac ako jednu loptičku. Uvedomili sme si, že umiestnenie loptičky predstavuje deň narodenín žiaka. Preto pre hľadané p máme

$$p = 1 - q = 1 - \frac{365\,364\,363\,...\,336}{365^{30}} = 1 - \frac{365}{365}\,\frac{364}{365}\,\frac{363}{365}\cdots\frac{336}{365}$$

Situácia prekvapuje numerickými hodnotami, p je totiž neočakávane veľké, $p = 0.7063$. Keď je v triede 40 žiakov, tak dostávame už $p = 0.8912$. Len pre zaujímavosť, uveďme ďalšie hodnoty (pravdepodobnosti v tabuľke sú zaokrúhlené na 4 desatinné miesta)

n	30	40	50	60	70	80	90
p	0.7063	0.8912	0.9704	0.9941	0.9992	0.9999	1.0000

1.2.15 (*Rozmiestnenie k nerozlíšiteľných obálok do n priečinkov so zákazom*) Pretože ide o rozmiestnenie so zákazom, priečinok môže obsahovať najviac jednu obálku – teda jednu, alebo žiadnu. Samozrejme, nutne musí platiť $k \leq n$ (inak sa rozmiestnenie nedá realizovať). Rozmiestnenie je určené k prvkovou podmnožinou množiny $\{1, 2, \ldots, n\}$. Prvky tej podmnožiny sú čísla práve tých priečinkov, ktoré obsahujú obálku. Preto (podľa odstavca 1.2.8) počet možných rozmiestnení sa rovná $C(n, k)$.

1.2.16 Príklad. Majme 10 priečinkov (očíslovaných od 1 po 10) a 8 nerozlíšiteľných obálok. Hovorme o náhodnom rozmiestnení so zákazom. Aká je pravdepodobnosť toho, že

a) prvý priečinok bude prázdny?
b) prvý a druhý priečinok bude prázdny?
c) indexy prázdnych priečinkov sú párne?

Riešenie. Počet všetkých rozmiestnení sa rovná $C(10, 8)$, t. j. 45. Počet tých, ktoré realizujú to, že prvý priečinok je prázdny, sa rovná $C(9, 8)$. Preto pravdepodobnosť toho, že prvý priečinok neobsahuje obálku sa rovná $C(9, 8)/C(10, 8) = 9/45 = 0.2$.

Počet rozmiestnení, ktoré realizujú udalosť spočívajúcu v tom, že prvý a druhý priečinok je prázdny sa rovná 1, pretože je k dispozícii 8 priečinkov a 8 obálok musí byť rozmiestnených so zákazom. Preto pravdepodobnosť v časti (b) sa rovná 1/45.

Pre udalosť v časti (c) sa počet priaznivých rozmiestnení rovná $C(5, 2) = 10$ (5 priečinkov má párne indexy). Preto pravdepodobnosť v časti (c) sa rovná 10/45.

Úlohy

1.2.1 Náhodný pokus spočíva v hode štyrmi riadnymi, ale označenými mincami. Určte pravdepodobnosti náhodných udalostí:

a) na druhej minci padne znak,
b) znak padne len na druhej minci,
c) na prvej a na tretej minci padne znak,
d) znak padne len na prvej a tretej minci,
e) padnú práve dva znaky.

1.2.2 V škatuli máme 50 výrobkov a medzi nimi 5 nepodarkov. Realizujme náhodný výber bez vrátenia rozsahu 10. Aká je pravdepodobnosť toho, že medzi vybratými

a) nie je nepodarok?
b) je jeden nepodarok?
c) sú dva nepodarky?
d) sú najviac dva nepodarky?

1.2.3 V škatuli máme 100 výrobkov a medzi nimi 10 nepodarkov. Náhodne vyberieme 20 (bez vrátenia) a zistíme, že medzi vybratými je jeden nepodarok. S akou pravdepodobnosťou medzi ďalšími desiatimi náhodne vybratými (opäť bez vrátenia)

a) nie je nepodarok?
b) je jeden nepodarok?
c) sú dva nepodarky?
d) sú najviac dva nepodarky?

1.2.4 Zámok na heslo má na spoločnej osi päť kotúčikov a na obvode každého z nich sú cifry 0, 1, 2, 3, 4, 5, 6. Zámok sa otvára len na jedinú päticu cifier (heslo zámku). Aká je pravdepodobnosť otvorenia zámku, ak

a) náhodne zvolíme nejakú päticu cifier?
b) vieme, že v hesle sa žiadna cifra neopakuje?
c) vieme, že v hesle sa nevyskytuje nula?

1.2.5 V škatuli máme 5 bielych, 4 červené a 2 modré guľky. Náhodne vyberáme naraz dve. Aká je pravdepodobnosť toho, že

a) obe budú biele?
b) vytiahneme bielu a červenú?
c) vytiahnuté guľky budú guľky rovnakej farby?

1.2.6 V škatuli máme 5 bielych, 4 červené a 2 modré guľky. Náhodne vyberáme naraz tri. Aká je pravdepodobnosť toho, že

a) sú rovnakej farby?
b) medzi nimi nie je biela?
c) medzi vybratými je aspoň jedna modrá?
d) vytiahnuté guličky budú guličky len dvoch farieb?

1.2.7 V škatuli je 9 lístkov očíslovaných číslami 1, 2, 3, ..., 9. Náhodne vyberieme naraz dva lístky. Aká je pravdepodobnosť toho, že

a) čísla na vytiahnutých lístkoch sa líšia o viac ako o dva?
b) súčet čísel na vytiahnutých lístkoch je aspoň 7?
c) súčin čísel na vytiahnutých lístkoch je aspoň 9?

1.2.8 V škatuli je 9 lístkov očíslovaných číslami 1, 2, 3, ..., 9. Náhodne vyberieme dva lístky tak, že najprv ťaháme prvý, poznačíme si číslo na ňom, vrátime ho späť a potom ťaháme druhý. Aká je pravdepodobnosť toho, že

a) čísla na vytiahnutých lístkoch sa líšia o viac ako o dva?
b) súčet čísel na vytiahnutých lístkoch je aspoň 7?
c) súčin čísel na vytiahnutých lístkoch je aspoň 9?

1.2.9 Pokus spočíva v hádzaní piatimi mincami, ktorých ruby, resp. líca sme označili číslami 1 resp. 2. Aká je pravdepodobnosť toho, že súčet padnutých čísel je 7 alebo 8?

1.2.10 Z balíčka 52 francúzskych kariet sme si náhodne vytiahli päť kariet. Aká je pravdepodobnosť toho, že

a) máme v ruke 4 karty rovnakej hodnoty?
b) máme v ruke práve dvoch kráľov?
c) máme v ruke práve tri karty rovnakej hodnoty?
d) máme v ruke práve tri karty rovnakej farby?

Terminologická poznámka: balíček má karty štyroch farieb – trefy, kára, srdcia a piky.

1.2.11 V škatuli máme 9 lístkov očíslovaných od 1 po 9. Náhodne vyberajme jeden po druhom, (dodržujúc poradie vytiahnutých), pričom vytiahnuté nedávame späť. Vytiahnutím posledného lístku dostávame náhodnú permutáciu číslic 1 až 9. Aká je pravdepodobnosť toho, že

a) v náhodnej permutácii číslice 3 a 7 budú na svojich miestach?
b) v náhodnej permutácii číslice 1, 2, 3, 4 budú na svojich miestach?

Všimnime si, že niečo iné je pýtať sa na udalosť „práve číslice 3, 7 budú na svojich miestach“, resp. na udalosť „práve číslice 1, 2, 3, 4 budú na svojich miestach“ (to sú ťažšie otázky).

c) nebude existovať „zhoda“ t. j. žiadna číslica nebude na svojom mieste? (Táto otázka má motivovať, budeme sa jej venovať neskôr.)

1.2.12 Do šiestich (prázdnych) skladov rozdeľme náhodne 30 výrobkov takým spôsobom, aby všetkých 6^{30} rozmiestnení bolo rovnako možných. Aká je pravdepodobnosť toho, že

a) prvý sklad ostane prázdny?
b) prvý sklad bude obsahovať práve 5 výrobkov?
c) prvý sklad bude obsahovať 5 a druhý 7 výrobkov?

1.2.13 V škatuli máme n výrobkov a medzi nimi m nepodarkov $(0 < m < n)$. Realizujme neusporiadaný náhodný výber bez vrátenia rozsahu k. Aká je pravdepodobnosť toho, že medzi vybranými je i nepodarkov, keď pre i platí: $\max(0, m+k-n) \le i \le \min(m, k)$? Všimnite si, že ide o zovšeobecnenie úlohy 1.2.2, kde $n = 50$, $m = 5$, $k = 10$.

1.2.14 Náhodne rozmiestnime n rozlíšiteľných obálok do n priečinkov tak, aby všetkých n^n rozmiestnení bolo rovnako pravdepodobných (pozri odsek 1.2.12). Aká je pravdepodobnosť toho, že

a) prvý priečinok ostane prázdny?
b) iba prvý priečinok ostane prázdny?

1.2.15 V škatuli máme n výrobkov očíslovaných od 1 po n. Uvažujme o neusporiadanom výbere bez vrátenia rozsahu k. Označme

A_i – výrobok s číslom i sa dostane do výberu,
B_{ij} – výrobok s číslom i a aj výrobok s číslom j sa dostane do výberu,
C_{ij} – výrobok s číslom i sa dostane, ale výrobok s číslom j sa nedostane do výberu.

Určte $P(A_i)$, $P(B_{ij})$, $P(C_{ij})$.

1.3 Diskrétny pravdepodobnostný priestor

Z článku 1.1 vieme, že model náhodného pokusu je trojica symbolov: (Ω, S, P). Ω predstavuje množinu všetkých potenciálne možných výsledkov pokusu, S je systém podmnožín množiny Ω a modeluje náhodné udalosti. Povedať, že A je náhodná udalosť, je to isté, ako napísať $A \in S$. Symbol P predstavuje pravdepodobnosť. Pritom P chápeme ako zobrazenie, ktoré udalostiam A (t. j. prvkom systému S) priraďuje ich pravdepodobnosť $P(A)$. Zdôrazňujeme, že

- jedna vec je pre udalosti konkrétneho pokusu vedieť stanoviť ich pravdepodobnosti, teda pre $A \in S$ vedieť stanoviť *hodnotu* $P(A)$ tak, aby model správne popisoval skutočnosť a
- druhá vec je – poznať vlastnosti, ktoré má *každá* pravdepodobnosť.

Prvá vec je náročnejšia. Aj preto, že nie je vždy jednoduché postrehnúť, že v našom modelovaní je niečo zlé, že niečo nesedí. Čo sa týka druhej veci, teda vlastností pravdepodobnosti vo všeobecnosti, tú prácu už urobili iní. Budeme si všímať, čo o tom hovoria vety tejto kapitoly.

V tomto článku bude množina Ω alebo konečná, alebo ak nekonečná, tak spočítateľná. Za systém S môžeme v týchto prípadoch vziať systém všetkých podmnožín množiny Ω. Aké vlastnosti má (každá) pravdepodobnosť P na systéme S? Motiváciou pre nasledujúcu definíciu sú vlastnosti tých konkrétnych pravdepodobností, ktoré sme rozoberali v článkoch 1.1 a 1.2. Získali sme ich buď uplatnením klasickej Laplaceovej definície, alebo pravdepodobnosti sme chápali ako limity postupností relatívnych početností. V oboch prípadoch – keď P chápeme ako zobrazenie – je pravdepodobnosť P *aditívna*. V definícii 1.3.1 použijme zápis $P(\omega_i)$ namiesto presného $P(\{\omega_i\})$ a toto zjednodušenie budeme používať aj ďalej.

1.3.1 Definícia. Pravdepodobnostný priestor (Ω, S, P) sa nazýva *diskrétny*, ak Ω je konečná, t. j. $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$, alebo nekonečná, ale spočítateľná, t. j. $\Omega = \{\omega_1, \omega_2, \dots, \omega_n, \dots\}$, S je systém všetkých podmnožín množiny Ω, $P\colon S \to \langle 0, 1\rangle$ je zobrazenie s vlastnosťami:

1. $\sum_i P(\omega_i) = 1$,
2. pre všetky $A \in S$ platí $P(A) = \sum_{i:\omega_i \in A} P(\omega_i)$.

Ak $\Omega = \{\omega_1, \omega_2, \dots, \omega_n, \dots\}$, tak súčet $\sum_i P(\omega_i)$ je súčtom nekonečného radu. Analogicky, ak udalosť A má nekonečne veľa prvkov, $\sum_{i:\omega_i \in A} P(\omega_i)$ je tiež súčet nekonečného radu. Často použijeme označenie $p_i = P(\omega_i)$. Ak napr. $A = \{\omega_1, \omega_3, \omega_7, \omega_9\}$, tak $P(A) = \sum_{i:\omega_i \in A} P(\omega_i)$ znamená, že $P(A) = p_1 + p_3 + p_7 + p_9$.

Všimnime si, že vlastnosť (2) zaručuje, že P je *aditívna*. Vskutku, ak A a B sú disjunktné, teda ak $A \cap B = \emptyset$, tak $\{i\colon \omega_i \in A\} \cap \{j\colon \omega_j \in B\} = \emptyset$, a preto

$$P(A \cup B) = \sum_{k:\,\omega_k \in A\cup B} P(\omega_k) = \sum_{i:\,\omega_i \in A} P(\omega_i) + \sum_{j:\,\omega_j \in B} P(\omega_j) = P(A) + P(B).$$

Vlastnosti (1), (2) znamenajú, že pravdepodobnosť (v prípade diskrétneho priestoru) je daná svojimi hodnotami p_i na elementárnych udalostiach $\{\omega_i\}$, teda na výsledkoch pokusu.

Hodnoty $p_i = \mathrm{P}(\omega_i)$ určujú hodnoty P(A) pre akúkoľvek udalosť A. Takto v prípade diskrétneho priestoru namiesto trojice $(\Omega, \mathcal{S}, \mathrm{P})$, za model pokusu môžeme vziať jednoducho dvojicu vektorov:

$$\underline{\omega} = (\omega_1, \omega_2, \dots, \omega_n), \quad \text{prípadne,} \quad \underline{\omega} = (\omega_1, \omega_2, \dots, \omega_n, \dots)$$
$$\underline{p} = (p_1, p_2, \dots, p_n), \quad \underline{p} = (p_1, p_2, \dots, p_n, \dots)$$

Ak je Ω konečná a ak $p_i = \mathrm{P}(\omega_i) = 1/n$, pre $i = 1, 2, \dots, n$, hovoríme, že priestor je *homogénny*. Používa sa aj označenie *klasický* pravdepodobnostný priestor. Hovorili sme o ňom v čl. 1.2.

1.3.2 Veta. Ak $(\Omega, \mathcal{S}, \mathrm{P})$ je diskrétny pravdepodobnostný priestor, tak pre $\mathrm{A}, \mathrm{B} \in \mathcal{S}$ platí:

a) P je aditívna, t. j. ak $\mathrm{A} \cap \mathrm{B} = \emptyset$, tak $\mathrm{P}(\mathrm{A} \cup \mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B})$.

b) Ak $\mathrm{A} \subset \mathrm{B}$ potom $\mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B} \setminus \mathrm{A}) = \mathrm{P}(\mathrm{B})$, a preto $\mathrm{P}(\mathrm{A}) \le \mathrm{P}(\mathrm{B})$.

c) $\mathrm{P}(\mathrm{A}') = 1 - \mathrm{P}(\mathrm{A})$.

d) $\mathrm{P}(\mathrm{A} \cup \mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B} \setminus \mathrm{A}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B} \setminus (\mathrm{A} \cap \mathrm{B}))$.

e) $\mathrm{P}(\mathrm{A} \cup \mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B}) - \mathrm{P}(\mathrm{A} \cap \mathrm{B})$.

Dôkaz. Aditivitu P sme dokázali v odseku 1.3.1. Všetky ďalšie body sú dôsledkom aditivity. V bode (b) využijeme: ak $\mathrm{A} \subset \mathrm{B}$, potom $\mathrm{A} \cup (\mathrm{B} \setminus \mathrm{A}) = \mathrm{B}$, a preto

$$\mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B} \setminus \mathrm{A}) = \mathrm{P}(\mathrm{B}), \text{ odkiaľ } 0 \le \mathrm{P}(\mathrm{B} \setminus \mathrm{A}) = \mathrm{P}(\mathrm{B}) - \mathrm{P}(\mathrm{A}), \text{ teda } \mathrm{P}(\mathrm{A}) \le \mathrm{P}(\mathrm{B}).$$

Bod (c) je dôsledok triviálneho faktu: $1 = \mathrm{P}(\Omega) = \mathrm{P}(\mathrm{A} \cup \mathrm{A}') = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{A}')$.

V bode (d) uplatníme vzťahy medzi udalosťami (teda medzi množinami):

$$\mathrm{A} \cup \mathrm{B} = \mathrm{A} \cup (\mathrm{B} \setminus \mathrm{A}) = \mathrm{A} \cup (\mathrm{B} \setminus (\mathrm{A} \cap \mathrm{B})).$$

Pretože A a $(\mathrm{B} \setminus \mathrm{A})$ sú disjunktné, resp. A a $(\mathrm{B} \setminus (\mathrm{A} \cap \mathrm{B}))$ sú disjunktné, to, čo chceme dokázať, vyplýva z aditivity. Dôkaz bodu (e) môžeme oprieť o bod (d) a bod (b).

Podľa (d) máme $\mathrm{P}(\mathrm{A} \cup \mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B} \setminus (\mathrm{A} \cap \mathrm{B}))$. Pretože $(\mathrm{A} \cap \mathrm{B}) \subset \mathrm{B}$, podľa bodu (b) platí

$$\mathrm{P}(\mathrm{A} \cap \mathrm{B}) + \mathrm{P}(\mathrm{B} \setminus (\mathrm{A} \cap \mathrm{B})) = \mathrm{P}(\mathrm{B}), \text{ odkiaľ } \mathrm{P}(\mathrm{B} \setminus (\mathrm{A} \cap \mathrm{B})) = \mathrm{P}(\mathrm{B}) - \mathrm{P}(\mathrm{A} \cap \mathrm{B}).$$

Nakoniec,

$$\mathrm{P}(\mathrm{A} \cup \mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B} \setminus (\mathrm{A} \cap \mathrm{B})) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B}) - \mathrm{P}(\mathrm{A} \cap \mathrm{B}).$$

1.3.3 Príklad. Hádžeme dvoma riadnymi hracími kockami. Za výsledok považujeme súčet padnutých bodov. Aký je model tohoto náhodného pokusu? Aká je pravdepodobnosť toho, že súčet bodov (ktoré padnú na kockách) bude menší ako 8?

Riešenie. Teraz máme 11 možných výsledkov: $\Omega = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ a priestor nie je homogénny, veď kým výsledok $\omega_1 = 2$ nastáva iba ak na oboch kockách padne jeden bod, tak výsledok $\omega_6 = 7$ realizuje 6 možností: (1, 6) (2, 5) (3, 4) (4, 3) (5, 2) (6, 1). Pri stanovení $\mathrm{P}(\omega_i)$ si pomôžeme predstavou prvotného (pomocného) Ω_0, ktorý tvorí 36 rovnako pravdepodobných bodov priestoru Ω_0 (s Ω_0 sme pracovali v príklade 1.1.4). Úvahou, akou sme určili $\mathrm{P}(\omega_1)$, resp. $\mathrm{P}(\omega_6)$, nájdeme pravdepodobnosti všetkých jedenástich výsledkov:

ω_i	2	3	4	5	6	7	8	9	10	11	12
$\mathrm{P}(\omega_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

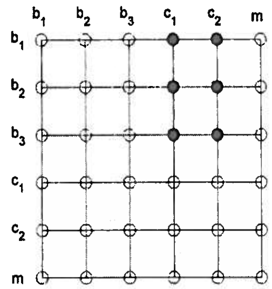
Pri stanovení $\mathrm{P}(\omega_i)$ sme využili predstavu pomocného homogénneho priestoru Ω_0 a aditivitu. Teraz je už jednoduché nájsť pravdepodobnosť toho, že súčet bodov bude menší ako 8. Zrejme ide o udalosť $\mathrm{A} = \{2, 3, 4, 5, 6, 7\}$, a preto

$$\mathrm{P}(\{2, 3, 4, 5, 6, 7\}) = \mathrm{P}(\mathrm{A}) = \sum_{i:\omega_i \in A} \mathrm{P}(\omega_i) = \frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} + \frac{5}{36} + \frac{6}{36} = \frac{21}{36} = \frac{7}{12}$$

1.3.4 Príklad. V škatuli máme 3 biele, 2 červené a 1 modrú loptičku. Náhodne vytiahneme prvú, poznačíme si jej farbu, vrátime ju do škatule a náhodne ťaháme druhú. Za výsledok považujeme usporiadanú dvojicu vytiahnutých farieb. Určime model tohoto náhodného pokusu.

Riešenie. Pretože výsledkom je usporiadaná dvojica farieb, Ω má 9 prvkov:

$$\Omega = \{(b, b), (b, c), (b, m), (c, b), (c, c), (c, m), (m, b), (m, c), (m, m)\} = \{\omega_1, \omega_2, \dots, \omega_9\}$$



Intuitívne je zrejmé, že $\mathrm{P}(\omega_1) > \mathrm{P}(\omega_9)$, teda náš priestor nie je homogénny.

$\mathrm{P}(\omega_i)$ nájdeme analogicky ako v 1.3.3. Pomocným homogénnym priestorom Ω_0 bude (zhodou okolností) opäť 36-bodový priestor rovnako pravdepodobných výsledkov (obr. vľavo). Vznikol tak, že loptičky si pomyslíme ako očíslované a uvažujeme o šiestich objektoch b_1, b_2, b_3, c_1, c_2, m. Je zrejmé, že vytiahnutie ľubovoľného z nich je rovnako pravdepodobné. Stačí si uvedomiť, koľko výsledkov pomocného priestoru Ω_0 vytvára jednotlivé body priestoru Ω.

Napr. 9 uzlov v pomocnom priestore Ω_0 (sú to uzly v prvých troch riadkoch a v prvých troch stĺpcoch) odpovedá výsledku (b, b). Na obrázku je zvýraznených tých 6 bodov pomocného Ω_0, ktoré vytvárajú výsledok (b, c).

Za model pokusu môžeme vziať dva vektory

vektor výsledkov: $(\omega_1, \omega_2, \dots, \omega_9)$

a vektor pravdepodobností: $(p_1, p_2, \dots, p_9)$.

ω_i	(b, b)	(b, c)	(b, m)	(c, b)	(c, c)	(c, m)	(m, b)	(m, c)	(m, m)
p_i	9/36	6/36	3/36	6/36	4/36	2/36	3/36	2/36	1/36

1.3.5 Príklad. Škatuľa a jej obsah z predchádzajúceho príkladu ostávajú, ale réžia pokusu sa zmení. Teraz prvú vytiahnutú nevrátime späť pred ťahaním druhej. Za výsledok považujeme opäť usporiadanú dvojicu vytiahnutých farieb. Určime model pokusu.

Riešenie. Takýto pokus je, samozrejme, iný ako v 1.3.4, veď výsledok (m, m) nie je teraz možný. Opäť ide o nehomogénny priestor. Model pokusu predstavuje tabuľka:

ω_i	(b, b)	(b, c)	(b, m)	(c, b)	(c, c)	(c, m)	(m, b)	(m, c)
p_i	6/30	6/30	3/30	6/30	2/30	2/30	3/30	2/30

Ako sme našli pravdepodobnosti? Výsledok pokusu určujú výsledky dvoch akcií: prvou je ťahanie prvej – tu je 6 možností (v škatuli je 6 objektov). Druhou akciou je druhý ťah, pričom v škatuli je už len 5 objektov (prvá vytiahnutá sa nevrátila do škatule). Bodov pomocného priestoru je preto 30 (= 6·5). Tých 30 rôznych usporiadaných dvojíc, ktoré môžu byť vytiahnuté, predstavuje 30 bodov pomocného priestoru Ω_0, ktorý je homogénny.

Napríklad, výsledku $\omega_1 = (b, b)$ odpovedá 6 bodov pomocného (homogénneho) priestoru Ω_0: (b_1, b_2), (b_1, b_3), (b_2, b_1), (b_2, b_3), (b_3, b_1), (b_3, b_2). Alebo výsledku $\omega_2 = (b, c)$ odpovedá týchto šesť bodov priestoru Ω_0: (b_1, c_1), (b_1, c_2), (b_2, c_1), (b_2, c_2), (b_3, c_1), (b_3, c_2). Analogicky získame p_3 až p_8. Vždy je dobré overiť, že súčet všetkých p_i sa rovná jednej. To síce nie je záruka toho, že sme $P(\omega_i)$ stanovili správne, ale ak súčet sa nerovná 1, tak je isté, že je v našom modelovaní chyba.

1.3.6 Príklad. Škatuľa a jej obsah z predchádzajúceho príkladu ostávajú, ale teraz náhodne ťaháme naraz dve loptičky, a preto je prirodzené za výsledok pokusu považovať neusporiadanú dvojicu vytiahnutých farieb. Určime model pokusu.

Riešenie. Model pokusu predstavuje tabuľka:

ω_i	*b-b*	*b-c*	*b-m*	*c-c*	*c-m*
p_i	3/15	6/15	3/15	1/15	2/15

Teraz je výsledkom neusporiadaná dvojica farieb. Farby sú tri, ale objektov je 6. Keď použijeme opäť pomocné indexovanie, tak je zrejmé, že kým k výsledku (*c-c*) vedie len jedna dvojica červených (sú tam len dve červené), tak k výsledku (*b-c*) vedie 6 prvotných výsledkov: $\{b_1, c_1\}$, $\{b_1, c_2\}$, $\{b_2, c_1\}$, $\{b_2, c_2\}$, $\{b_3, c_1\}$, $\{b_3, c_2\}$. Tentoraz prvotné výsledky chápeme ako neusporiadané dvojice. Pomocným homogénnym priestorom je priestor všetkých možných neusporiadaných dvojíc. Zrejme má 15 (= C(6, 2)) bodov.

1.3.7 Príklad. Náhodný pokus spočíva v hádzaní mincou dovtedy, kým nepadne znak (akonáhle na minci padne znak, pokus skončil). Navrhnime priestor možných výsledkov tak, aby sme mohli modelovať udalosti

a) A – v pokuse sa bude hádzať aspoň tri razy,
b) B – počet hodov v pokuse bude párne číslo.

Určime pravdepodobnoti jednotlivých výsledkov a nájdime P(A), resp. P(B).

Riešenie. Teraz množina možných výsledkov pokusu je nekonečná (ale spočítateľná), pretože

$$\Omega = \{\omega_1, \omega_2, \omega_3, \dots, \omega_n, \dots\} = \{(Z), (C, Z), (C, C, Z), (C, C, C, Z), \dots, (C, C, \dots, C, Z), \dots\}.$$

Pre udalosti máme

$$A = \{\omega_3, \omega_4, \omega_5, \dots, \omega_n, \dots\} = \{(C, C, Z), (C, C, C, Z), \dots, (C, C, \dots, C, Z), \dots\},$$

$$B = \{\omega_2, \omega_4, \omega_6, \dots, \omega_{2k}, \dots\} = \{(C, Z), (C, C, C, Z), (C, C, C, C, C, Z), \dots\}.$$

O tom, že $P(\omega_1) = \frac{1}{2}$, asi nikto nepochybuje. Avšak to, že $P(\omega_i) = \frac{1}{2^i}$, pre $i = 2, 3, \dots, n, \dots$ nemusí byť také samozrejmé (to objasníme neskôr).

Všimnime si, že takéto modelovanie pravdepodobnosti nás nedovedie k sporu, pretože zrejme platí $\sum_{i=1}^{\infty} P(\omega_i) = 1$. P(A), resp. P(B), nájdeme uplatnením definície 1.3.1.

$$P(A) = \sum_{i:\omega_i\in A} P(\omega_i) = \frac{1}{2^3} + \frac{1}{2^4} + \frac{1}{2^5} + \cdots = \frac{1}{2^3}\left(1 + \frac{1}{2} + \frac{1}{2^2} + \cdots\right) = \frac{1}{2^3}(1 + 1) = \frac{1}{4}$$

$$P(B) = \sum_{i:\omega_i\in B} P(\omega_i) = \frac{1}{2^2} + \frac{1}{2^4} + \frac{1}{2^6} + \cdots = \frac{1}{2^2}\left(1 + \frac{1}{2^2} + \frac{1}{2^4} + \cdots\right) = \frac{1}{2^2}\left(1 + \frac{1}{3}\right) = \frac{1}{3}$$

P(A) aj P(B) sme určili ako súčet geometrického radu. Samozrejme, P(A) je možné nájsť aj jednoduchšie, pretože určiť P(A′) je ľahké. Opačnou udalosťou k A je udalosť spočívajúca v tom, že v pokuse sa bude hádzať raz, alebo dva razy, preto P(A′) = 0.5 + 0.25 = 0.75 (a to je v zhode s tým, že P(A) = 0.25).

Je možné namietať, že medzi výsledky pokusu sme mali zahrnúť aj ten, keď sa padnutia znaku nedočkáme. Taký výsledok, označili by sme ho napr. ω_0, by predstavovala nekonečná postupnosť (C, C, ..., C, ...). V takom prípade by do množiny možných výsledkov pribudol bod ω_0. Položili by sme však $P(\omega_0) = 0$, a tak by sa na modelovaní nášho pokusu prakticky nič nezmenilo. Jeden z argumentov na podporu toho, že $P(\omega_0) = 0$, sa opiera o vlastnosť spojitosti pravdepodobnosti, ktorú formulujeme v nasledujúcom článku.

Úlohy

1.3.1 Náhodný pokus spočíva v hode troma hracími kockami. Nech A je udalosť – súčet bodov, ktoré padli, sa rovná 11 a B nech je udalosť – súčet bodov, ktoré padli, sa rovná 12. Ktorá z týchto udalostí je pravdepodobnejšia?

1.3.2 Pokus spočíva v hode šiestimi hracími kockami. Nech A je udalosť – na kockách padlo 6 rôznych čísel, B je udalosť – na každej kocke padlo párne číslo. Ktorá z týchto udalostí je pravdepodobnejšia?

1.3.3 Náhodný pokus spočíva v hádzaní mincou dovtedy, kým nepadne znak dva razy po sebe. Navrhnite priestor možných výsledkov a určte pravdepodobnosti udalostí

a) A – v pokuse sa bude hádzať aspoň štyri razy,
b) B – v pokuse sa nebude hádzať viac ako desaťkrát.

1.3.4 Hádžeme hracou kockou dovtedy, kým nepadne šestka. Navrhnite priestor možných výsledkov tak, aby bolo možné modelovať nasledujúce udalosti

a) A – pokus končí tretím hodom, pričom v každom hode padnú aspoň štyri body,
b) B – pokus končí štvrtým hodom, a pritom v každom hode padne párne číslo,
c) C – počet hodov v pokuse bude nepárny.

1.4 Kolmogorovov model náhodného pokusu

Doteraz išlo o náhodné pokusy, v ktorých množina Ω všetkých možných výsledkov bola konečná, resp. spočítateľná. V takých prípadoch sme si mohli dovoliť dve veci:

- Systém náhodných udalostí modelovať systémom všetkých podmnožín množiny Ω.
- Pravdepodobnosť udalostí definovať pomocou pravdepodobností jednotlivých výsledkov pokusu:

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

Obe skutočnosti sme si ľahko osvojili. Teraz však musíme zobrať na vedomie fakt, že ak Ω nie je spočítateľná, tak z istých dôvodov (jemnej matematickej povahy) sa s takýmito predpokladmi pracovať nedá.

Namiesto systému všetkých podmnožín je nutné pracovať so sigma algebrami podmnožín množiny Ω. Tiež je pravda, že pravdepodobnosť nie je vždy možné definovať cez pravdepodobnosť elementárnych udalostí (napr. aj preto, lebo často pravdepodobnosť každého jednotlivého výsledku sa rovná nule).

Pred sformulovaním všeobecnej definície pravdepodobnostného priestoru, rozoberme dve jednoduché situácie, v ktorých je množina Ω možných výsledkov nespočítateľná.

1.4.1 Príklad. (Náhodný bod x intervalu $\langle a,b\rangle$) Nech $a,b\in\mathrm{R}$, $a<b$ a uvažujme o tom, že na interval $\langle a,b\rangle$ „hádžeme" bod x tak, že sú splnené tieto predpoklady:

- Je isté, že v každom hode bodom x vždy zasiahneme interval, teda vždy $x\in\langle a,b\rangle$.
- Pravdepodobnosť toho, že $x\in\langle c,d\rangle$, kde $\langle c,d\rangle$ je podinterval intervalu $\langle a,b\rangle$, sa rovná podielu $\frac{d-c}{b-a}$, to znamená, že

$$\mathrm{P}(\, x \text{ padne do } \langle c,d\rangle\,) = \mathrm{P}(\, c\le x\le d\,) = \frac{d-c}{b-a}$$

Pre zjednodušenie, nech $a=0$, $b=1$, teda bod x hádžeme na interval $\langle 0,1\rangle$. V takom prípade $\mathrm{P}(c\le x\le d)=d-c$. Napr. $\mathrm{P}(0.2\le x\le 0.5)=0.3$, analogicky, $\mathrm{P}(0.6\le x\le 0.8)=0.2$, a preto (odvolávajúc sa na aditivitu P) máme

$$\mathrm{P}(\, x\in\langle 0.2,0.5\rangle\cup\langle 0.6,0.8\rangle\,) = 0.3+0.2=0.5$$

Týmto spôsobom pravdepodobnosť priradíme aj zjednoteniam disjunktných intervalov, a preto aj komplementom takých množín. Dá sa ukázať (ale to je nad naše sily), že nie je možné rozšíriť takúto pravdepodobnosť na všetky podmnožiny intervalu $\langle 0,1\rangle$. Avšak rozšíriť našu pravdepodobnosť z intervalov na veľmi bohatý systém podmnožín možné je. Všimnime si dve skutočnosti:

- Pravdepodobnosť P je *rovnomerne* rozložená na $\langle 0,1\rangle$ v tom zmysle, že ak $\langle c,d\rangle$ je podinterval $\langle 0,1\rangle$, tak pravdepodobnosť $\mathrm{P}(c\le x\le d)$ nezávisí od polohy intervalu $\langle c,d\rangle$ v $\langle 0,1\rangle$. To znamená, že ak $\langle c+h,d+h\rangle$ je tiež podinterval $\langle 0,1\rangle$, tak platí

$$\mathrm{P}(c+h\le x\le d+h) = \mathrm{P}(c\le x\le d) = d-c,$$

  pretože podľa definície: $\mathrm{P}(c+h\le x\le d+h) = d+h-(c+h) = d-c = \mathrm{P}(c\le x\le d)$.

- Pre každé $u\in(0,1)$ máme

$$\mathrm{P}(x=u)=0,$$

  pretože $\mathrm{P}(x=u)\le \mathrm{P}(u\le x\le u+1/n) = u+1/n-u = 1/n$ (ak n je dostatočne veľké, aby $u+1/n\le 1$), z čoho, ak $n\to\infty$ máme: $\mathrm{P}(x=u)=0$.

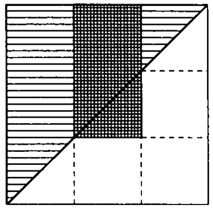
1.4.2 Príklad. Spolužiačky Danka a Janka prichádzajú do triedy každé ráno vždy v intervale $\langle 7{:}40, 7{:}55\rangle$. To je interval dĺžky 15 (minút). Zjednodušme to a hovorme, že príchod Danky, resp. Janky, predstavuje náhodný bod x, resp. náhodný bod y, pričom oba body x, y sú body intervalu $\langle 0,15\rangle$. Za výsledok pokusu berme bod (x,y) štvorca $\langle 0,15\rangle\times\langle 0,15\rangle$.

Situácia je analógiou príkladu 1.4.1. Tam išlo o podintervaly $\langle c,d\rangle$ intervalu $\langle a,b\rangle$, teraz pôjde o dvojrozmerné intervaly štvorca $\Omega=\langle 0,15\rangle\times\langle 0,15\rangle$. Uvažujme o udalostiach:

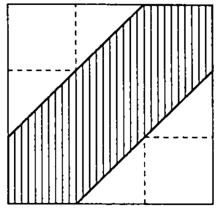
$\mathrm{A}=\{(x,y): 5\le x<10,\ 5\le y<15\}$, t. j. Danka príde medzi 7:45 a 7:50, Janka medzi 7:45 a 7:55

$\mathrm{B}=\{(x,y): x\le y\}$, teda B spočíva v tom, že Danka príde skôr, alebo spolu s Jankou,
$\mathrm{C}=\{(x,y): |x-y|\le 5\}$, C spočíva v tom, že jedna nečaká na druhú viac ako 5 minút.

Udalosti modelujeme ako podmnožiny štvorca $\langle 0,15\rangle\times\langle 0,15\rangle$. Kým udalosť A je obdĺžnik, B je trojuholník (obr. 1-2a). Udalosť C je zvýraznená na obr. 1-2b.



Obr. 1-2a



Obr. 1-2b

Nech bod (x,y) je *náhodný* výsledok pokusu a (u,v) je ľubovoľný bod štvorca. Z analogických dôvodov ako v 1.4.1 platí

$$\mathrm{P}(x=u,\ y=v)=0,$$

čo znamená, že pravdepodobnosť nie je možné definovať prostredníctvom výsledkov. Môžeme ju ale celkom jednoducho definovať pre špeciálne udalosti – dvojrozmerné intervaly.

Ak $\mathrm{A}=\langle a,b\rangle\times\langle c,d\rangle$ je podmnožinou Ω, tak pravdepodobnosť toho, že náhodný bod padne do A definujeme vzťahom:

$$\mathrm{P}(\,(x,y)\in\mathrm{A}\,) = \frac{\text{plošný obsah } A}{\text{plošný obsah } \Omega} = \frac{(b-a)(d-c)}{225}$$

Preto máme $\mathrm{P(A)}=\frac{50}{225}$. P(B) stanovíme ako plochu trojuholníka, $\mathrm{P(B)}=\frac{\frac{1}{2}15.15}{225}=0.5$, resp. $\mathrm{P(C)}=\frac{5}{9}$. Takto stanovenej pravdepodobnosti hovoríme *geometrická pravdepodobnosť*. Situácia ilustruje aj to, že niekedy nedokážeme prisúdiť pravdepodobnosť každej podmnožine množiny Ω (pretože nevieme stanoviť plošný obsah *každej* podmnožiny štvorca).

V 30. rokoch 20. storočia A. N. Kolmogorov navrhol axiomatickú výstavbu teórie pravdepodobnosti. Axiomatický prístup nedáva návod ako pre danú A nájsť P(A), ale spočíva v tom, že formuluje tie vlastnosti, ktoré má pravdepodobnosť v každom pokuse, pričom na pravdepodobnosť nazeráme ako na zobrazenie. Kým Kolmogorov použil vlastnosť sigma-aditivity P, v našej definícii 1.4.3 použijeme spojitosť. Je to idea jednoduchšia a pritom (v kontexte aditivity P) sú oba pojmy (teda sigma-aditivita a spojitosť) ekvivalentné.

1.4.3 Definícia. *Pravdepodobnostný priestor* je trojica (Ω, S, P), kde

Ω je neprázdna množina (body $\omega\in\Omega$ interpretujeme ako výsledky náhodného pokusu),
S je sigma algebra podmnožín množiny Ω, t. j. S je systém podmnožín s vlastnosťami

a) $\Omega\in S$,
b) ak $\mathrm{A}\in S$, tak $\mathrm{A}'\in S$,
c) ak $\mathrm{A}_n\in S$, pre $n=1,2,3,\dots$, tak $\bigcup_{n=1}^{\infty}\mathrm{A}_n\in S$.

P je pravdepodobnosť, t. j. zobrazenie, P: $S \to \langle 0, 1 \rangle$, pre ktoré platí

1. $P(\emptyset) = 0$, $P(\Omega) = 1$,
2. ak $A, B \in S$, $A \cap B = \emptyset$, tak $P(A \cup B) = P(A) + P(B)$, (*aditivita* P)
3. ak $A_n \in S$, $A_n \subset A_{n+1}$, $n = 1, 2, 3, \ldots$, tak $P(\bigcup_{n=1}^{\infty} A_n) = \lim_{n\to\infty} P(A_n)$ (*spojitosť* P)

1.4.4 Veta (Ďalšie vlastnosti pravdepodobnosti) Ak (Ω, S, P) je pravdepodobnostný priestor, tak pre $A, B, C \in S$ platí:

a) $P(A') = 1 - P(A)$.

b) Ak A, B sú také, že $A \subset B$, tak $P(A) \le P(B)$ a $P(B) = P(A) + P(B \setminus A)$.

c) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

d) $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$.

e) P je konečne aditívna: Ak $A_i \in S$, $i = 1, 2, \ldots, n$, pričom $A_i \cap A_j = \emptyset$, pre $i \ne j$, tak

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i).$$

f) P je subaditívna: Ak $A_i \in S$, $i = 1, 2, \ldots, n$, tak $P(\bigcup_{i=1}^{n} A_i) \le \sum_{i=1}^{n} P(A_i)$.

V dôkaze vety 1.3.2 sme ukázali, že body (a), (b), (c) sú dôsledky aditivity P. Dôkaz bodu (d) je obsah príkladu 1.4.5. Body (e), (f) sú dôsledkom prvých dvoch definitorických vlastností P. Dá sa ukázať, že ak P má vlastnosti (1) a (2), tak vlastnosť (3), je ekvivalentná vlastnosti, ktorej sa hovorí *sigma-aditivita* P:

$$\text{Ak } A_n \in S,\ n = 1, 2, \ldots,\ A_i \cap A_j = \emptyset,\ (i \ne j),\ i, j = 1, 2, \ldots,\ \text{tak } P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n).$$

Je zrejmé, že sigma-aditivita je zosilnením konečnej aditivity na nekonečne veľa udalostí.

1.4.5 Príklad. Dôkaz bodu (d) sa dá urobiť rôznymi spôsobmi. Samozrejme, všetky sa opierajú o aditivitu. Jeden zo spôsobov je využiť prácu s atómami, generovanými udalosťami A, B, C. Čo sú atómy? Najprv zaveďme označenie:
Nech A^0 znamená A a nech A^1 znamená A' (komplement A), analogicky pre B, resp. C.

Atóm je udalosť typu $A^i \cap B^j \cap C^k$, kde $i, j, k \in \{0, 1\}$. Napríklad, $A^0 \cap B^1 \cap C^0 = A \cap B' \cap C$, alebo $A^1 \cap B^1 \cap C^0 = A' \cap B' \cap C$ sú atómy.

Najprv si uvedomme, že rôzne atómy sú disjunktné. Koľko rôznych atómov generuje trojica udalostí A, B, C? Zrejme 8, pretože atóm je daný trojicou exponentov a exponentom je alebo 0, alebo 1. Počet trojíc exponentov sa rovná 8 ($= 2^3$). Pomocou atómov dokážeme opísať nielen A, B, C, ale aj $A \cap B$, $A \cap C$ a $B \cap C$. Napr.

$$A = (A^0 \cap B^1 \cap C^1) \cup (A^0 \cap B^0 \cap C^1) \cup (A^0 \cap B^1 \cap C^0) \cup (A^0 \cap B^0 \cap C^0),$$

$$\text{resp. } A \cap B = (A^0 \cap B^0 \cap C^0) \cup (A^0 \cap B^0 \cap C^1).$$

Dôkaz bodu (d) urobíme takto: $A \cup B \cup C$ je zjednotenie siedmych atómov $M_1, \ldots, M_7$, ten ôsmy, ktorý sa na úvahe nezúčastňuje, je $A^1 \cap B^1 \cap C^1$. Pretože atómy sú disjunktné, platí $P(A \cup B \cup C) = P(M_1) + \ldots + P(M_7)$.

Ukážme, že pravá strana vzťahu (d) sa tiež rovná tomuto súčtu. Napr. $P(A) - P(A \cap B)$ sa rovná $P(A^0 \cap B^1 \cap C^1) + P(A^0 \cap B^1 \cap C^0)$ a analogicky pre ďalšie členy. K dokončeniu dôkazu stačí overiť, že súčet všetkých členov pravej strany sa tiež rovná $P(M_1) + \ldots + P(M_7)$.

1.4.6 Príklad. Objasnime, že ak $A_i \in S$, $i = 1, 2, \ldots, n$, tak platí:

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) - \sum_{\substack{i,j=1\\ i<j}}^{n} P(A_i \cap A_j) + \sum_{\substack{i,j,k=1\\ i<j<k}}^{n} P(A_i \cap A_j \cap A_k) - \ldots +$$
$$+ (-1)^{n-1} P(A_1 \cap A_2 \cap \ldots \cap A_n)$$

Formálny dôkaz sa urobí indukciou ku n. Pre $n = 1$ nie je čo dokazovať, pre $n = 2$ ide o známy vzťah, je to bod (c) vety 1.4.4. Pre $n = 3$ sme v 1.4.5 dôkaz naznačili použitím atómov. V druhom kroku (dôkazu indukciou) využijeme aditivitu P. Nebudeme ho však robiť. Skôr nás zaujíma, kedy takýto vzťah môžeme s úspechom použiť. O tom hovorí nasledujúci (určite zaujímavý) príklad.

1.4.7 Príklad. Uvažujme o náhodnej permutácii dĺžky n. Položme si otázky:

1. Čo je pravdepodobnejšie? To, že zhoda v permutácii existuje, alebo to, že neexistuje?
2. Závisí odpoveď na prvú otázku od dĺžky permutácie?

Riešenie. O náhodnej permutácii sme hovorili v 1.2.6 a 1.2.7. Nech A_i znamená udalosť – na mieste i je zhoda (pozor, A_i nehovorí nič iné, len to, že na mieste i je zhoda). Potom $A_i \cap A_j$ je udalosť – na miestach i, j sú zhody (pozor, to neznamená, že v permutácii sú dve zhody, len to, že zhody sú na miestach i, j), analogicky $A_i \cap A_j \cap A_k$, ... atď.

Udalosť „existuje zhoda" je udalosť $A_1 \cup A_2 \cup \ldots \cup A_n$ a vypočítať $P(A_1 \cup A_2 \cup \ldots \cup A_n)$ pomocou vzťahu z 1.4.6 nie je problém, pretože určiť $P(A_i)$ je ľahké, rovnako ľahké je určiť $P(A_i \cap A_j)$, $P(A_i \cap A_j \cap A_k)$, ... atď. Zrejme platí (pozri príklad 1.2.7)

$$P(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n}, \quad P(A_i \cap A_j) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}, \quad P(A_i \cap A_j \cap A_k) = \frac{(n-3)!}{n!} = \frac{1}{n(n-1)(n-2)} \quad \text{atď.}$$

Takto máme

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = n\,\frac{1}{n} - \binom{n}{2}\frac{1}{n(n-1)} + \binom{n}{3}\frac{1}{n(n-1)(n-2)} - \cdots + (-1)^{n-1}\frac{1}{n!} =$$
$$= 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \cdots + (-1)^{n-1}\frac{1}{n!}$$

Pravdepodobnosť toho, že v n-tici existuje zhoda je stanovená. Pre pravdepodobnosť opačnej udalosti máme

$$P(\text{neexistuje zhoda}) = 1 - P(\text{existuje zhoda}) = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \cdots + (-1)^n \frac{1}{n!} \approx e^{-1}.$$

Ak $n = 9$ (ako v 1.2.7), tak namiesto presného $\frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \ldots + (-1)^9 \frac{1}{9!} = 0.367\,879\,188$ (zaokrúhlené na 9 platných miest), môžeme vziať $e^{-1} = 0.367\,879\,441$ a vidíme, že chyba aproximácie je menšia ako $0.26\,10^{-6}$.

Ak teda ide o náhodnú permutáciu dĺžky 9, tak pravdepodobnosť toho, že neexistuje zhoda, sa rovná (približne) 0.368. To znamená, že je pravdepodobnejšie, že zhoda existuje. Je prekvapujúce, že odpoveď na prvú otázku temer nezávisí od dĺžky permutácie!

1.4.8 Úloha. Uvažujme o náhodnom rozmiestnení (bez zákazu) k rozlíšiteľných obálok do n priečinkov. Aká je pravdepodobnosť toho, že aspoň jeden priečinok je prázdny?

Návod. Nech A_i znamená – priečinok i je prázdny. Chceme nájsť $P(A_1 \cup A_2 \cup \ldots \cup A_n)$. Uvedomme si, že nie je ťažké určiť $P(A_i)$, $P(A_i \cap A_j)$, ... atď. a ďalej postupovať ako v 1.4.7.

1.5 Podmienená pravdepodobnosť

Často sa v rámci náhodného pokusu pýtame na pravdepodobnosť udalosti A, pritom vďaka okolnostiam vieme, že v pokuse nastala udalosť B. Tá skutočnosť (teda ten fakt, že v pokuse B nastala) niekedy viac a niekedy menej podmieňuje (ovplyvňuje) pravdepodobnosť udalosti A. Už teda nejde o pôvodnú P(A), ale o „novú“, podmienenú $P_B(A)$. Hovoríme, že udalosť B je *podmieňujúcou* udalosťou a odrazu je na scéne okrem pôvodnej pravdepodobnosti P(.) aj pravdepodobnosť nová, podmienená $P_B(.)$.

1.5.1 Príklad. V škatuli máme 20 lístkov s číslami od 1 po 20. Pokus spočíva v náhodnom ťahaní lístku zo škatule. Uvažujme o udalostiach
A – číslo na vytiahnutom lístku je nanajvýš 15,
B – číslo na vytiahnutom lístku je aspoň 11.

Je zrejmé, že pre pôvodné (nepodmienené) pravdepodobnosti platí: $P(A) = \frac{15}{20} = \frac{3}{4}$, resp. $P(B) = \frac{10}{20} = \frac{1}{2}$. Teraz pozmeňme úlohu. Predstavme si, že niekto už uskutočnil pokus a poskytol informáciu o tom, že v pokuse udalosť B nastala. Po tejto informácii máme odpovedať na otázku: Aká je pravdepodobnosť toho, že nastala A?

Zrejme teraz nejde o P(A), ale o $P_B(A)$ a uvažujeme asi takto: Nevieme síce aký lístok bol vytiahnutý, ale musel to byť lístok s číslom aspoň 11: 11, 12, 13, ... , 19, 20. To je 10 možností. Pre udalosť A je priaznivých 5 prípadov, preto

$$P_B(A) = \frac{5}{10} = \frac{1}{2}$$

Ako vidíme, pôvodná, nepodmienená P(A) = 0.75 a podmienená $P_B(A) = 0.50$. Podmieňovanie zmenšilo pravdepodobnosť nastatia A. Všimnime si, že pre vzťah medzi pôvodnou P a novou podmienenou P_B platí

$$P_B(A) = \frac{5}{10} = \frac{\frac{5}{20}}{\frac{10}{20}} = \frac{P(A\cap B)}{P(B)}$$

1.5.2 Definícia. Nech (Ω, *S*, P) je pravdepodobnostný priestor a nech B je udalosť s kladnou pravdepodobnosťou. Pravdepodobnosť *podmienenú* udalosťou B, nazývame pravdepodobnosť P_B definovanú pre všetky $A \in S$ vzťahom

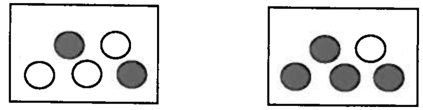
$$P_B(A) = \frac{P(A\cap B)}{P(B)}$$

Často sa definitorický vzťah pre $P_B(A)$ použije na výpočet $P(A\cap B)$ podľa vzťahu

$$P(A\cap B) = P(B)\,P_B(A), \quad \text{alebo} \quad P(A\cap B) = P(A)\,P_A(B).$$

Druhý vzťah sme získali len zámenou úloh udalostí A, B (veď podmieňujúcou môže byť aj udalosť A). Podmienená pravdepodobnosť teda slúži na výpočet $P(A\cap B)$ vtedy, ak podmienená pravdepodobnosť $P_B(A)$ je buď zadaná, alebo ak v danej situácii je jednoduché stanoviť ju bez definitorického vzťahu. Takú situáciu ilustruje príklad 1.5.3.

1.5.3 Príklad. Majme dve škatule, v prvej (ľavej) sú 3 biele a 2 čierne loptičky, v druhej (pravej) jedna biela a 4 čierne. Pokus spočíva v náhodnom ťahaní loptičky z ľavej škatule a jej vloženi do pravej a v následnom ťahaní loptičky z pravej škatule. S akou pravdepodobnosťou
a) sú obe vytiahnuté loptičky biele?
b) prvá je biela a druhá je čierna?



Riešenie. Nech B_1, resp. B_2 znamená prvá (resp. druhá) vytiahnutá je biela, analogicky udalosti C_1, C_2 sa týkajú čiernej. V tejto situácii využijeme definitorický vzťah v podobe

$$P(A\cap B) = P(A)\,P_A(B),$$

teda na výpočet pravdepodobnosti *prieniku* udalostí. Podmienenú pravdepodobnosť teraz vyčíslime ako klasickú pravdepodobnosť, keď vezmeme do úvahy, že pomery v pravej škatuli sú nové. Konkrétne

$$P(B_1\cap B_2) = P(B_1)P_{B_1}(B_2) = \frac{3}{5}\,\frac{2}{6} = \frac{1}{5},$$

pretože ak prvá vytiahnutá loptička je biela, tak zloženie pravej škatule pred ťahaním druhej je: 2 biele a 4 čierne. Analogicky

$$P(B_1\cap C_2) = P(B_1)P_{B_1}(C_2) = \frac{3}{5}\,\frac{4}{6} = \frac{2}{5}.$$

1.5.4 Poznámka. Keď podmieňujeme udalosťou, ktorá je vyjadrená ako prienik, alebo zjednotenie nejakých udalostí, je naše označenie nepraktické. Napríklad je nepraktické písať $P_{B\cap C\cap D}(A)$. Preto namiesto $P_B(A)$ sa vžilo označenie $P(A\,|\,B)$. Avšak, pozor, toto nové označenie trochu zavádza – treba mať na zreteli, že udalosťou, o pravdepodobnosť ktorej ide, je len a len udalosť A. Napríklad, namiesto $P_{B\cap C\cap D}(A)$ budeme písať $P(A\,|\,B\cap C\cap D)$, čo znamená, že cieľom je nájsť pravdepodobnosť A, keď podmieňujeme udalosťou $B\cap C\cap D$.

1.5.5 Veta. Nech (Ω, *S*, P) je pravdepodobnostný priestor, A, B, C ∈ *S*, pričom $P(A\cap B) > 0$. Potom

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

Dôkaz. Pracujme s pravou stranou a za $P(B\,|\,A)$, aj za $P(C\,|\,A\cap B)$ dosaďme podľa definitorického vzťahu. Členy v menovateľoch sa vykrátia a dostávame $P(A\cap B\cap C)$. Rôvnosť je dokázaná. Pritom (rovnakým postupom) sa ukáže aj všeobecnejšie:

Ak $A_1, A_2, \ldots, A_n \in S$ a sú také, že $P(A_1\cap A_2\cap\ldots\cap A_{n-1}) > 0$, potom

$$P(A_1\cap A_2\cap\ldots\cap A_n) = P(A_1)\,P(A_2\,|\,A_1)\,P(A_3\,|\,A_1\cap A_2)\ldots P(A_n\,|\,A_1\cap A_2\cap\ldots\cap A_{n-1}).$$

1.5.6 Príklad. Pridajme k dvom škatuliam z príkladu 1.5.3 tretiu škatuľu, v ktorej sú dve biele a dve čierne loptičky. Pokus spočíva v náhodnom ťahaní loptičky z prvej škatule a jej vloženi do druhej a v následnom ťahaní loptičky z druhej škatule a vloženi do tretej a konečne, náhodnom ťahaní loptičky z tretej škatule. S akou pravdepodobnosťou
a) sú všetky tri vytiahnuté loptičky biele?
b) je prvá a tretia vytiahnutá biele a druhá čierna?

Riešenie. Označme B_1, resp. B_2, resp. B_3 udalosti – prvá je biela, resp. druhá je biela, resp. tretia je biela. Analogicky pre čierne, napríklad, C_2 bude udalosť – druhá je čierna. Potom

a) $P(B_1 \cap B_2 \cap B_3) = P(B_1)P(B_2 | B_1)P(B_3 | B_1 \cap B_2) = P(B_1)P(B_2 | B_1)P(B_3 | B_2) = \frac{3}{25}$

b) $P(B_1 \cap C_2 \cap B_3) = P(B_1)P(C_2 | B_1)P(B_3 | B_1 \cap C_2) = P(B_1)P(C_2 | B_1)P(B_3 | C_2) = \frac{4}{25}$

V bode (a) sme použili rovnosť $P(B_3 | B_1 \cap B_2) = P(B_3 | B_2)$, ktorá, pochopiteľne, vo všeobecnosti neplatí. Avšak teraz – v našej situácii – platí preto, lebo pre pravdepodobnosť vytiahnutia bielej z tretej škatule je rozhodujúce iba to, akej farby loptička putovala z druhej škatule do tretej. Analogicky v bode (b) sme použili rovnosť $P(B_3 | B_1 \cap C_2) = P(B_3 | C_2)$.

1.5.7 Veta. (*O úplnej pravdepodobnosti*) Nech (Ω, S, P) je pravdepodobnostný priestor a $H_1, H_2, \dots, H_k \in S$ sú rozkladom Ω (to znamená, že $H_i \cap H_j = \emptyset$, pre $i \neq j$, $i, j = 1, 2, \dots, k$, a pritom $H_1 \cup H_2 \cup \dots \cup H_k = \Omega$). Nech $P(H_i) > 0$, pre $i = 1, 2, \dots, k$. Potom pre každú $A \in S$ platí

$$P(A) = \sum_{i=1}^{k} P(H_i)P(A|H_i).$$

Dôkaz využíva iba aditivitu P (v podobe konečnej aditivity, bod (e) z vety 1.4.4) a definitorický vzťah pre podmienenú pravdepodobnosť.

$$P(A) = P(A \cap (\cup_{i=1}^{k} H_i)) = P(\cup_{i=1}^{k}(A \cap H_i)) = \sum_{i=1}^{k} P(A \cap H_i) = \sum_{i=1}^{k} P(H_i)P(A|H_i),$$

pretože systém udalostí $\{A \cap H_i\}$, $i = 1, 2, \dots, k$, je systém disjunktných udalostí.

1.5.8 Príklad. Uvažujme dve škatule z príkladu 1.5.3, teda ľavá obsahuje 3 biele a 2 čierne, pravá jednu bielu a štyri čierne loptičky. Náhodný pokus spočíva v náhodnej voľbe škatule a náhodnom ťahaní loptičky z nej. K voľbe škatule použijeme hraciu kocku. Ak na kocke padne menej ako 5, volíme ľavú škatuľu, inak pravú. S akou pravdepodobnosťou bude

a) vytiahnutá loptička biela a bude ťahaná z ľavej škatule?

b) vytiahnutá loptička čierna a bude ťahaná z pravej škatule?

c) vytiahnutá loptička biela?

Riešenie. Označme H_1 – loptičku ťaháme z ľavej škatule, H_2 – loptičku ťaháme z pravej škatule. Nech B je udalosť – vytiahnutá je biela, resp. C – vytiahnutá je čierna. Na zodpovedanie prvých dvoch otázok nič nové nepotrebujeme. Teda postupujeme ako doteraz.

a) $P(B \cap H_1) = P(H_1 \cap B) = P(H_1)P(B | H_1) = \frac{4}{6}\frac{3}{5} = \frac{2}{5}$

b) $P(C \cap H_2) = P(H_2 \cap C) = P(H_2)P(C | H_2) = \frac{2}{6}\frac{4}{5} = \frac{4}{15}$

c) V tomto bode uplatníme aj aditivitu. Loptička totiž musela byť vytiahnutá z ľavej, alebo z pravej škatule, t. j. $H_1 \cup H_2 = \Omega$ (v hre sú dve hypotézy). Zrejme pre udalosti platí:

$$B = \Omega \cap B = (H_1 \cup H_2) \cap B = (H_1 \cap B) \cup (H_2 \cap B)$$

a vďaka aditivite

$$P(B) = P(H_1 \cap B) + P(H_2 \cap B) = P(H_1)P(B | H_1) + P(H_2)P(B | H_2) = \frac{4}{6}\frac{3}{5} + \frac{2}{6}\frac{1}{5} = \frac{7}{15}$$

Vidíme, že náš postup viedol ku vzťahu, ktorý deklaruje veta 1.5.7 (pre $k = 2$).

1.5.9 Veta. (Bayes) Nech platia predpoklady vety 1.5.7, nech $A \in S$ a $P(A) > 0$. Potom

$$P(H_i|A) = \frac{P(H_i \cap A)}{P(A)} = \frac{P(H_i)P(A|H_i)}{\sum_{j=1}^{k} P(H_j)P(A|H_j)}$$

1.5.10 Príklad. Uvažujme o náhodnom pokuse z príkladu 1.5.3. Predpokladajme, že loptičku vytiahnutú z ľavej škatule dávame do pravej bez toho, aby sme si všimli jej farbu. Ak loptička vytiahnutá z pravej škatule je biela, s akou pravdepodobnosťou loptička vytiahnutá z ľavej škatule bola biela?

Riešenie. Nevieme nič o farbe prvej ťahanej loptičky, ale sú tu len dve možnosti (hypotézy) H_1 – loptička ťahaná z ľavej škatule je biela, H_2 – loptička ťahaná z ľavej škatule je čierna. Ďalej nech B je udalosť – loptička vytiahnutá z pravej škatule je biela. Naša úloha je určiť $P(H_1 | B)$. Použijeme Bayesovu vetu, ale, ako vidíme, potrebujeme $P(B)$. Podľa vety 1.5.7 máme

$$P(B) = P((H_1 \cup H_2) \cap B) = P((H_1 \cap B) \cup (H_2 \cap B)) = P(H_1)P(B | H_1) + P(H_2)P(B | H_2) =$$
$$= \frac{3}{5}\frac{2}{6} + \frac{2}{5}\frac{1}{6} = \frac{4}{15}$$

Nakoniec

$$P(H_1|B) = \frac{P(H_1 \cap B)}{P(B)} = \frac{P(H_1)P(B|H_1)}{P(B)} = \frac{\frac{3}{5}\frac{2}{6}}{\frac{4}{15}} = \frac{15}{20} = \frac{3}{4}$$

Vidíme, že kým $P(H_1) = 0.6$, tak $P(H_1 | B) = 0.75$. Hovoríme, že *apriórna* pravdepodobnosť $P(H_1)$ sa zmenila na *aposteriórnu* $P(H_1 | B)$, ktorej hodnota sa rovná 0.75. Skrátka, akési udalosti – hypotézy H_i – majú pred pokusom svoje pravdepodobnosti $P(H_i)$, hovoríme im *apriórne pravdepodobnosti*. Po uskutočnení pokusu, keď vieme, že nastala udalosť B, sa tieto apriórne $P(H_i)$ zmenia na *aposteriórne* pravdepodobnosti $P(H_i | B)$.

1.5.11 Poznámka. Nech (Ω, S, P) je pravdepodobnostný priestor, nech $C \in S$ a $P(C) > 0$. Podmienenú pravdepodobnosť $P(. | C)$ akceptujeme ako „novú" pravdepodobnosť, ktorá reflektuje to, že pôvodné podmienky pokusu sa zmenili. Pritom podmienená pravdepodobnosť $P(. | C)$ má všetky atribúty pojmu pravdepodobnosť z definície 1.4.3. Teda $P(. | C)$ môžeme chápať ako zobrazenie z $S \to \langle 0, 1 \rangle$, ktoré (okrem iného) je aditívne (presvedčte sa)

ak $A, B \in S$ a $A \cap B = \emptyset$, tak $P(A \cup B | C) = P(A | C) + P(B | C)$.

Úlohy

1.5.1 Uvažujme náhodný pokus spočívajúci v hode dvoma obyčajnými hracími kockami, modrou a červenou. Označme náhodné udalosti takto:

A na oboch kockách rovnaký počet bodov,

B na modrej kocke párny počet bodov,

C na červenej aspoň 5 bodov,

D súčet bodov na kockách je najviac 6.

Stanovte a porovnajte:

a) $P(A)$ a $P(A | D)$, overte, že $P(A) < P(A | D)$.

b) $P(B)$ a $P(B | D)$, tentoraz $P(B) > P(B | D)$.

c) $P(B)$ a $P(B | C)$, teraz $P(B) = P(B | C)$.

d) $P(C)$ a $P(C | D)$, $P(C | D)$ je podstatne menšia ako $P(C)$.

1.5.2 Náhodný pokus spočíva v hode tromi označenými hracími kockami. Označme

A súčet bodov na všetkých kockách sa rovná 10,

B maximum z bodov na prvých dvoch kockách sa rovná 6.

Nájdite $P(A)$, $P(B)$, $P(A \mid B)$, $P(B \mid A)$.

1.5.3 Pridajme k dvom škatuliam z príkladu 1.5.3 tretiu škatuľu, v ktorej sú dve biele a dve čierne loptičky. Pokus spočíva v náhodnom ťahaní loptičky z prvej škatule a jej vložení do druhej a v následnom ťahaní loptičky z druhej a vložení do tretej a konečne náhodnom ťahaní loptičky z tretej škatule. S akou pravdepodobnosťou

a) sú všetky tri vytiahnuté loptičky čierne?
b) je prvá a tretia vytiahnutá čierna a druhá biela?
c) je prvá a tretia vytiahnutá čierna?
d) tretia vytiahnutá je čierna?

1.5.4 Uvažujme tri škatule z predchádzajúcej úlohy, ale teraz loptičku vytiahnutú z tretej škatule vložíme do prvej škatule. S akou pravdepodobnosťou bude zloženie prvej škatule také, ako bolo na začiatku? Teda prvá bude obsahovať opäť 3 biele a dve čierne loptičky.

1.5.5 Majme opäť naše dve škatule, v prvej (ľavej) 3 biele a 2 čierne loptičky, v druhej (pravej) jednu bielu a 4 čierne. Pokus spočíva v náhodnom ťahaní loptičky z ľavej škatule a jej vložení do pravej bez toho, aby sme si pozreli jej farbu a v následnom ťahaní loptičky z pravej škatule. S akou pravdepodobnosťou

a) bola prvá vytiahnutá biela, ak vieme, že loptička ťahaná z druhej škatule bola biela?
b) bola prvá vytiahnutá čierna, ak vieme, že loptička ťahaná z druhej škatule bola biela?
c) bola prvá vytiahnutá biela, ak vieme, že loptička ťahaná z druhej škatule bola čierna?
d) bola prvá vytiahnutá čierna, ak vieme, že loptička ťahaná z druhej škatule bola čierna?

Otázka: Ktoré pravdepodobnosti musia dať v súčte hodnotu 1?

1.5.6 Na vysielači sa objavujú tri signály: bodka, čiarka a medzera s pravdepodobnosťami 0.5, 0.3, 0.2. V dôsledku porúch prenosového kanála sa vyslaný signál môže prijať chybne. Označme prijatie bodky, čiarky a medzery symbolmi: B, C, M a ich vyslatie symbolmi H_b, H_c, H_m. Dané sú pravdepodobnosti

$$\begin{array}{lll} P(B \mid H_b) = 0.6 & P(C \mid H_b) = 0.2 & P(M \mid H_b) = 0.2 \\ P(B \mid H_c) = 0.2 & P(C \mid H_c) = 0.7 & P(M \mid H_c) = 0.1 \\ P(B \mid H_m) = 0.2 & P(C \mid H_m) = 0 & P(M \mid H_m) = 0.8 \end{array}$$

a) Nájdite pravdepodobnosti, s akými sa na prijímači objavuje bodka, resp. čiarka, resp. medzera.
b) Ak bola prijatá bodka, s akou pravdepodobnosťou bola vyslaná bodka?
c) Ak bola prijatá čiarka, s akou pravdepodobnosťou bola vyslaná bodka?
d) Ak bola prijatá medzera, s akou pravdepodobnosťou bola vyslaná bodka?

Otázka: Musia dať tri pravdepodobnosti z bodu (a) v súčte hodnotu 1? Musia dať pravdepodobnosti z bodov (b), (c) a (d) v súčte hodnotu 1?

1.5.7 Produkcia je tvorená tromi automatickými linkami, ktoré sa podieľajú na celkovej produkcii (po rade) 50 %, 30 %, resp. 20 %. Nepodarkovosť jednotlivých liniek sa rovná (po rade) 2 %, 3 %, resp. 4 %.

a) Aká je nepodarkovosť celej produkcie?
b) Nech náhodne vybratý výrobok z celkovej produkcie je nepodarok. Aká je pravdepodobnosť toho, že bol vyrobený treťou linkou?

1.5.8 Študent odpovedá formou testu, v ktorom ku každej otázke má ponúknutých 7 alternatívnych odpovedí. Ak má danú otázku naštudovanú, vyberá s istotou správnu odpoveď. V opačnom prípade náhodne volí jednu z ponúkaných odpovedí. Predpokladajme, že test je zostavený tak pozorne, že pravdepodobnosť uhádnutia správnej odpovede sa rovná 1/7. Nakoniec predpokladajme, že študent naštudoval 70 % určenej látky.

a) Aká je pravdepodobnosť, že na náhodne zvolenú otázku študent správne odpovedá?
b) Ak študent dal správnu odpoveď, s akou pravdepodobnosťou nehádal?

1.5.9 Sérii 100 výrobkov prisudzujeme prvú alebo druhú kvalitu nasledujúcim postupom. Vyberieme náhodne 10 výrobkov (výberom bez vrátenia). Ak medzi vybratými nebude chybný, tak sérii prisúdime prvú kvalitu. Ak medzi vybratými bude viac ako jeden chybný, tak sérii prisúdime druhú kvalitu. Ak medzi vybratými bude jeden chybný, tak spomedzi ostávajúcich 90 výrobkov vyberieme náhodne ďalších 10 (opäť výberom bez vrátenia). Ak medzi nimi nebude nepodarok, tak pôvodnej sérii 100 výrobkov prisúdime prvú kvalitu – v inom prípade sérii prisúdime druhú kvalitu.

Predpokladajme, že séria obsahuje 5 chybných výrobkov. S akou pravdepodobnosťou sérii prisúdime prvú kvalitu? S akou pravdepodobnosťou jej prisúdime druhú kvalitu?

1.5.10 Diagnostické zariadenie registruje zmenu vo výrobnom procese – ak o zmenu skutočne ide – s pravdepodobnosťou 0.99. Na druhej strane s pravdepodobnosťou 0.05 hlási zmenu v prípadoch, keď o zmenu v skutočnosti nejde. Predpokladajme, že pravdepodobnosť zmeny vo výrobnom procese sa rovná 0.05.

a) Aká je pravdepodobnosť toho, že zariadenie hlási zmenu?
b) Predpokladajme, že zariadenie hlási zmenu. S akou pravdepodobnosťou ide skutočne o zmenu vo výrobnom procese?

1.5.11 Zariadenie môže pracovať v troch režimoch. Pravdepodobnosť toho, že pracuje v prvom, druhom, resp. treťom režime sa rovná (po rade) 0.60, 0.25, resp. 0.15. Pravdepodobnosť poruchy závisí od toho, v akom režime zariadenie funguje a rovná sa (po rade) 0.01, 0.02, resp. 0.06.

a) Aká je pravdepodobnosť poruchy počas doby T?
b) Predpokladajme, že počas doby T došlo k poruche. Aké sú pravdepodobnosti toho, že k poruche došlo v prvom, druhom, resp. treťom režime?

1.5.12 Krvný test indikuje prítomnosť vírusu v krvi – ak vírus je naozaj prítomný – s pravdepodobnosťou 0.99. Na druhej strane s pravdepodobnosťou 0.02 indikuje prítomnosť vírusu aj vtedy, keď vírus prítomný nie je. Predpokladajme, že p100 % populácie má vírus v svojej krvi. S akou pravdepodobnosťou má človek vírus, ak jeho test je pozitívny?

1.5.13 Šperkovnica má tri zásuvky. Nepoznáme obsah žiadnej konkrétnej zásuvky, ale vieme, že v jednej z tých troch zásuviek sú 2 strieborné mince, v ďalšej strieborná a zlatá a v tretej zásuvke sú 2 zlaté mince. Náhodne vyberieme zásuvku, z nej náhodne vyberieme mincu a zistíme, že je zlatá. Aká je pravdepodobnosť toho, že aj druhá minca v tej zásuvke je zlatá?

1.5.14 Pred nami je škatuľa, do ktorej nevidíme, ale vieme, že obsahuje 5 loptičiek. Ten, kto ich tam dal, nás ubezpečil, že obsahuje aspoň jednu bielu loptičku. Preto do úvahy prichádzajú tieto možnosti (hypotézy o obsahu škatule)

H_1 :	1 biela,	4 čierne	H_4 :	4 biele,	1 čierna
H_2 :	2 biele,	3 čierne	H_5 :	5 bielych,	0 čiernych
H_3 :	3 biele,	2 čierne			

Pretože nemáme žiadne ďalšie informácie (okrem tej, že škatuľa obsahuje aspoň jednu bielu), každej možnosti (každej hypotéze H_i) prisúdime rovnakú pravdepodobnosť. Hovoríme, že *apriórne* pravdepodobnosti sa rovnajú 0.2, t.j. pre každú H_i platí: $P(H_i) = 0.2$.

a) Predstavme si, že vykonáme pokus, v ktorom náhodne vytiahneme jednu loptičku. Predpokladajme, že bude čierna. Aké budú *aposteriórne* pravdepodobnosti hytopéz?

b) Predstavme si, že ešte raz (teda druhý raz) ťaháme loptičku (prvú sme naspäť nevrátili) a že bude opäť čierna. Aké budú teraz aposteriórne pravdepodobnosti hypotéz?

1.6 Nezávislosť udalostí

Nech modelom pokusu je priestor (Ω, S, P) a nech udalosti A, B majú nenulové pravdepodobnosti. Uvažujme o podmienených pravdepodobnostiach $P(A \mid B)$, resp. $P(B \mid A)$: Ak platí $P(A) = P(A \mid B)$, t.j. podmieňovanie udalosťou B nezmení pravdepodobnosť udalosti A, je prirodzené považovať udalosti A, B za *nezávislé*. Všimnime si, že platí:

Ak $P(A) = P(A \mid B)$, tak platí aj $P(A \cap B) = P(A)P(B)$, a preto aj $P(B) = P(B \mid A)$.

Pre definíciu nezávislosti udalostí sa berie vzťah $P(A \cap B) = P(A)P(B)$, ktorého prednosťou je, že je symetrický ku A, B a navyše, nevyžaduje predpokladať, že $P(A)$, $P(B)$ sú kladné.

1.6.1 Definícia. Nech A, B sú náhodné udalosti v priestore (Ω, S, P). Hovoríme, že A, B sú (stochasticky) nezávislé, ak platí $P(A \cap B) = P(A)P(B)$.

1.6.2 Príklad. Ukážme, že ak A, B sú nezávislé, tak nezávislé sú aj dvojice udalostí A a B', aj A', B a tiež A' a B'.
Riešenie. Ukážme, napríklad, že A, B' sú nezávislé, teda, že $P(A \cap B') = P(A)P(B')$. Zrejme

$$P(A \cap B') = P(A) - P(A \cap B) = P(A) - P(A)P(B) = P(A)(1 - P(B)) = P(A)P(B').$$

1.6.3 Príklad. Všimnime si, čo hovorí definícia nezávislosti pre situácie, v istom zmysle extrémne, napríklad:

a) Je istá udalosť Ω nezávislá s udalosťou A, pre akúkoľvek udalosť A? (Áno.)

b) Je nemožná udalosť $\emptyset$ nezávislá s udalosťou A, pre akúkoľvek udalosť A? (Áno.)

c) Môže byť A nezávislá s udalosťou A'? (Iba ak $P(A) = 0$, alebo $P(A) = 1$, inak nie.)

1.6.4 Príklad. Náhodný pokus spočíva v hode dvoma hracími kockami – modrou a červenou. Označme

A – na modrej menej ako 4 body,
B – na červenej menej ako 4 body,
C – súčet bodov na kockách sa rovná 7.

a) Overte, že každá dvojica udalostí (spomedzi A, B, C) je dvojica nezávislých udalostí.

b) Ukážte, že udalosť $A \cup B$ nie je nezávislá s udalosťou C.

c) Preverte, či platí $P(A \cap B \cap C) = P(A)P(B)P(C)$.

Riešenie.

a) Zrejme $P(A \cap B) = P(A)P(B)$, naviac, nezávislosť dvojice A, B je intuitívne zrejmá – A sa týka len modrej kocky a B len červenej. Aj dvojica A, C je dvojicou nezávislých udalostí, pretože platí $P(A \cap C) = P(A)P(C)$, keďže $P(A) = 1/2$, $P(C) = 1/6$ a $P(A \cap C) = 3/36 = 1/12$. Analogicky zistíme, že aj B, C je dvojica nezávislých udalostí.

b) Máme ukázať, že rovnosť $P((A \cup B) \cap C) = P(A \cup B)P(C)$ neplatí. Ľahko sa overí, že C je podudalosťou $A \cup B$, t.j. $C \subset A \cup B$. Preto ľavá strana sa rovná $P(C)$, kým pravá strana je menšia ako $P(C)$, teda rovnosť neplatí.

c) $A \cap B \cap C = \emptyset$, a preto $P(A \cap B \cap C) = 0$. Rovnosť $P(A \cap B \cap C) = P(A)P(B)P(C)$ neplatí, pretože $P(A)P(B)P(C) = 1/24$.

V príklade ide o trojicu udalostí, ktoré sú po dvoch nezávislé, ale nimi vykombinované udalosti nie sú nezávislé – veď udalosť $(A \cup B)$ nie je nezávislá s C. Ak chceme, aby $(A \cup B)$ bola nezávislá s C, musíme žiadať viac ako to, aby A, B, C boli po dvoch nezávislé. Preto sa nezávislosť trojice udalostí definuje tak, ako to uvádza nasledujúca definícia.

1.6.5 Definícia. Udalosti A, B, C sú (ako trojica) nezávislé, ak sú po dvoch nezávislé a naviac platí:

$$P(A \cap B \cap C) = P(A)P(B)P(C).$$

1.6.6 Poznámka. Uvedomme si, že predpoklad nezávislosti udalostí A, B, C znamená, že platia štyri rovnosti:

$$P(A \cap B) = P(A)P(B), \quad P(A \cap C) = P(A)P(C), \quad P(B \cap C) = P(B)P(C),$$
$$P(A \cap B \cap C) = P(A)P(B)P(C).$$

Ľahko sa ukáže, že ak A, B, C sú nezávislé udalosti, tak nezávislé sú aj dvojice, resp. trojice

a) A', B, C b) A', B', C c) A', B', C' d) $A \cap B$, C e) A, $B \cup C$
f) A, $B \cap C$ g) A', $B \cup C$ a pod.

Ukážme, napríklad, že A', $B \cup C$ sú nezávislé (bod g). Máme dokázať, že platí

$$P(A' \cap (B \cup C)) = P(A')P(B \cup C).$$

Zrejme

$$P(A' \cap (B \cup C)) = P((A' \cap B) \cup (A' \cap C)) = P(A' \cap B) + P(A' \cap C) - P(A' \cap B \cap C) = P(A')P(B) +$$
$$+ P(A')P(C) - P(A')P(B)P(C) = P(A')(P(B) + P(C) - P(B \cap C)) = P(A')P(B \cup C).$$

1.6.7 Definícia. Udalosti A, B, C, D sú (ako štvorica) nezávislé, ak každá trojica z nich je trojicou nezávislých udalostí a naviac platí

$$P(A \cap B \cap C \cap D) = P(A)\,P(B)\,P(C)\,P(D).$$

Slová „ako trojica", resp. „ako štvorica" budeme vynechávať. Neprehliadnime, že nezávislosť udalostí A, B, C, D znamená platnosť 6 + 4 + 1, t. j. 11 rovností. Totiž 6 rovností zaručí nezávislosť všetkých dvojíc, ďalšie 4 zaručia nezávislosť všetkých trojíc a jedenástou rovnosťou je rovnosť $P(A \cap B \cap C \cap D) = P(A)\,P(B)\,P(C)\,P(D)$.

Ukážte, že ak A, B, C, D sú nezávislé udalosti, tak nezávislými sú napr. aj

a) A', B, C, D b) A', B', C, D c) $A \cap B$, C, D d) $A \cup B$, $C \cap D$

Nech $A_1, A_2, \dots, A_n$ je ľubovoľná n-tica náhodných udalostí ($n \geq 2$, je ľubovoľné prirodzené číslo). Nasledujúca definícia formuluje nezávislosť n-tice náhodných udalostí a hoci po formálnej stránke je iná ako predchádzajúce, pre $n = 3$, resp. $n = 4$, sa obsahovo zhoduje s hore uvedenými formuláciami.

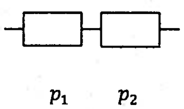
1.6.8 Definícia. Udalosti $A_1, A_2, \dots, A_n$ sú (ako n-tica) nezávislé, ak pre každú podmnožinu rôznych indexov $i_1, i_2, \dots, i_k$ takú, že $\{i_1, i_2, \dots, i_k\} \subset \{1, 2, \dots, n\}$, $2 \leq k \leq n$, platí

$$P\left(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}\right) = P(A_{i_1})\,P(A_{i_2}) \dots P(A_{i_k})$$

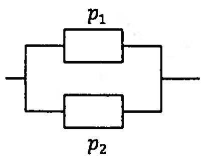
1.6.9 Poznámka. Spoľahlivosťou nejakého zariadenia, resp. jeho časti (hovorme jej blok) budeme rozumieť pravdepodobnosť jeho bezporuchovej práce počas nejakej pevne stanovenej doby T.

Schémy, ktoré používame v nasledujúcich úlohách treba chápať tak, že bloky v nich spájame „za sebou", ak zlyhanie ktoréhokoľvek z nich má za následok zlyhanie celého zariadenia. Naproti tomu spojenie blokov „vedľa seba" chápeme ako zdvojenie, resp. znásobenie daného bloku, t. j. zariadenie zlyhá z dôvodu zlyhania tohoto miesta práve vtedy, keď zlyhajú všetky bloky tohoto miesta zapojené „vedľa seba". Nasledujúce úlohy ukazujú, že výsledky predchádzajúcich úloh majú v praktických situáciách priame využitie. Všimnime si, že o uvažovaných blokoch predpokladáme, že sú nezávislé ako n-tice, teda v zmysle definície 1.6.8.

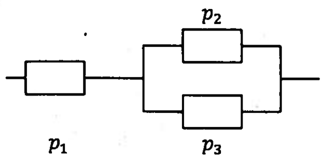
1.6.10 Príklad. Nech zariadenie pozostáva z blokov, ako je to naznačené na obrázkoch 1-3 a vysvetlené v 1.6.9. Nech spoľahlivosť jednotlivých blokov sa rovná p_1, p_2, p_3. Nájdite spoľahlivosť zariadenia v jednotlivých prípadoch za predpokladu, že javy zlyhania blokov sú nezávislé.



Obr. 1-3a.



Obr. 1-3b.



Obr. 1-3c.

Riešenie. Označme F udalosť, že zariadenie počas doby T funguje. Ak A_i znamená udalosť i-tý blok funguje, tak podľa predpokladu $P(A_i) = p_i$, pre $i = 1, 2, 3$. Ďalej

a) $F = A_1 \cap A_2$, $P(F) = P(A_1 \cap A_2) = P(A_1)\,P(A_2) = p_1 p_2$.

b) $F = A_1 \cup A_2$, $P(F) = P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) = p_1 + p_2 - p_1 p_2$.

c) $F = A_1 \cap (A_2 \cup A_3)$, $P(F) = P(A_1)\,P(A_2 \cup A_3)$, pretože A_1 je nezávislá s $A_2 \cup A_3$ podľa 1.6.6 e). Podľa (b) máme $P(A_2 \cup A_3) = p_2 + p_3 - p_2 p_3$ a nakoniec, $P(F) = p_1(p_2 + p_3 - p_2 p_3)$.

1.6.11 Príklad. Systém pozostáva z dvoch blokov typu I a troch blokov typu II. Náhoda ovplyvňuje fungovanie, resp. nefungovanie jednotlivých blokov. Označme udalosti takto:

A_i i-tý blok typu I funguje,
B_j j-tý blok typu II funguje.

Predpokladajme, že udalosti A_1, A_2, B_1, B_2, B_3 sú nezávislé (ako 5-tica, samozrejme) a platí: $P(A_i) = p$, $i = 1, 2$, $P(B_j) = q$, $j = 1, 2, 3$. Nájdite pravdepodobnosť

a) udalosti C, ktorá spočíva v tom, že funguje len druhý blok typu I a len tretí blok typu II,
b) udalosti D (núdzový režim), ktorá nastáva, ak funguje práve jeden blok typu I a súčasne práve jeden blok typu II,
c) udalosti E (spoľahlivý režim), ktorá znamená fungovanie aspoň jedného bloku typu I a súčasne fungovanie aspoň dvoch blokov typu II.

Riešenie. a) Zrejme $C = A_1' \cap A_2 \cap B_1' \cap B_2' \cap B_3$ a vďaka nezávislosti máme

$$P(C) = P(A_1' \cap A_2 \cap B_1' \cap B_2' \cap B_3) = (1-p)\,p\,(1-q)^2 q.$$

b) $D = (A_1 \cap A_2' \cap B_1 \cap B_2' \cap B_3') \cup (A_1 \cap A_2' \cap B_1' \cap B_2 \cap B_3') \cup (A_1 \cap A_2' \cap B_1' \cap B_2' \cap B_3) \cup$
$\cup\ (A_1' \cap A_2 \cap B_1 \cap B_2' \cap B_3') \cup (A_1' \cap A_2 \cap B_1' \cap B_2 \cap B_3') \cup (A_1' \cap A_2 \cap B_1' \cap B_2' \cap B_3)$,

teda D je zjednotením šiestich disjunktných udalostí, z ktorých každá má pravdepodobnosť rovnajúcu sa $(1-p)\,p\,(1-q)^2 q$, preto

$$P(D) = 6(1-p)\,p\,(1-q)^2 q.$$

c) $E = (A_1 \cup A_2) \cap [(B_1 \cap B_2) \cup (B_1 \cap B_3) \cup (B_2 \cap B_3)]$, pritom

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1)P(A_2) = p + p - p^2 = p(2-p),$$

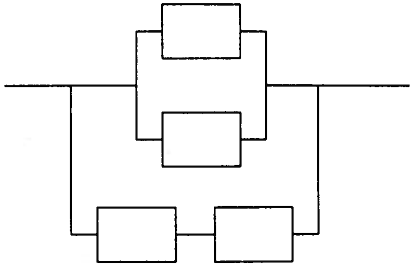
$$P((B_1 \cap B_2) \cup (B_1 \cap B_3) \cup (B_2 \cap B_3)) = P(B_1 \cap B_2) + P(B_1 \cap B_3) + P(B_2 \cap B_3) -$$
$$- P(B_1 \cap B_2 \cap B_3) - P(B_1 \cap B_2 \cap B_3) - P(B_1 \cap B_2 \cap B_3) + P(B_1 \cap B_2 \cap B_3) = 3q^2 - 2q^3.$$

To znamená, že

$$P(E) = P(A_1 \cup A_2)\,P((B_1 \cap B_2) \cup (B_1 \cap B_3) \cup (B_2 \cap B_3)) = p(2-p)q^2(3-2q).$$

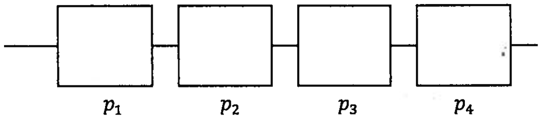
Úlohy

1.6.1 Systém pozostáva zo štyroch blokov (obr.1-4), ktorých spoľahlivosť (po rade zhora dole) sa rovná 0.7, 0.6, 0.8, 0.9. Vypočítajte spoľahlivosť systému, t. j. pravdepodobnosť nezlyhania systému, ak udalosti zlyhania blokov sú nezávislé.



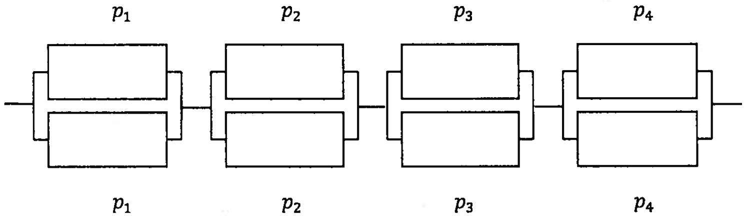
Obr. 1-4.

1.6.2 Systém pozostáva zo štyroch blokov podľa obr. 1-5 a spoľahlivosť jednotlivých blokov sa rovná p_1, p_2, p_3, p_4. Aká je spoľahlivosť takéhoto systému, ak je možné predpokladať nezávislosť zlyhania blokov?



Obr. 1-5.

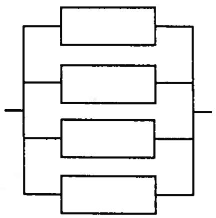
1.6.3 Pre zvýšenie spoľahlivosti systému z predchádzajúcej úlohy, zdvojme každý blok (pozri obr. 1-6). Ak opäť predpokladáme nezávislosť všetkých blokov, aká bude spoľahlivosť systému? Pre ilustráciu, porovnajte spoľahlivosť systému pred zdvojením a po zdvojení, pre konkrétne hodnoty $p_1 = p_2 = 0.9$, $p_3 = p_4 = 0.95$.



Obr. 1-6.

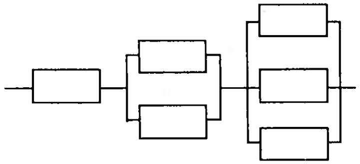
1.6.4 Uvažujme paralelný systém so štyrmi zložkami podľa obr. 1-7, ktorých spoľahlivosť označme p_1, p_2, p_3, p_4. Nech udalosti zlyhania sú nezávislé.

a) Vypočítajte spoľahlivosť systému.
b) Nech spoľahlivosť každej zložky sa rovná 0.8. Najmenej z koľkých takýchto zložiek musí takýto paralelný systém pozostávať, aby jeho spoľahlivosť bola aspoň 0.999?



Obr. 1-7.

1.6.5 Uvažujme o systéme na obr. 1-8, kde spoľahlivosti jednotlivých blokov sú zľava doprava po rade: 0.9, 0.8, 0.8 (pre dva bloky vedľa seba), 0.7, 0.7, 0.7 (pre tri bloky vedľa seba). Aká je spoľahlivosť celého systému, ak udalosti zlyhania blokov sú nezávislé?



Obr. 1-8.

1.6.6 Strelci Adam, Braňo a Cyril zasahujú terč s pravdepodobnosťami 0.6, 0.7 a 0.8, pričom udalosti zasahnutia terča Adamom, Braňom, resp. Cyrilom je trojica nezávislých udalostí. Predpokladajme, že vystrelia na (ten istý) terč súčasne. Aká je pravdepodobnosť toho, že

a) terč zasiahol iba Adam?
b) terč zasiahla len jedna strela?
c) terč zasiahli dve strely?
d) terč zasiahol Adam, ak vieme, že terč zasiahla len jedna strela?
e) terč zasiahol Adam, ak vieme, že terč zasiahli dve strely?

1.6.7 Kozmická sonda je vybavená tromi rovnakými na sebe nezávisle pracujúcimi kamerami. Kamery sú pre prípad poruchy vybavené korekčným zariadením. Pravdepodobnosť poruchy kamery sa rovná 0.2 a pravdepodobnosť opravy korekčným zariadením sa rovná 0.4. Aká je pravdepodobnosť, že

a) sonda nič neodfotografuje?
b) na sonde bude pracovať aspoň jedna kamera?
c) sonda odvysiela fotografie aspoň dvoch kamier?

1.6.8 Automat výstupnej kontroly označí chybný výrobok ako chybný s pravdepodobnosťou 0.99 a s pravdepodobnosťou 0.1 (čo je zrejme veľa) označí bezchybný ako chybný. Z toho dôvodu je do výstupnej kontroly zapojený druhý automat, ktorý označuje chybné ako chybné s pravdepodobnosťou 0.9 a dobré označí ako chybné s pravdepodobnosťou 0.05. Predpokladajme, že nepodarkovosť produkcie je $p100$ %-ná. Určte

a) percento tých, ktoré prejdú oboma automatmi ako bezchybné,
b) percento tých, ktoré prvý automat označí ako bezchybné a druhý ako chybné,
c) percento tých, ktoré prvý automat označí ako chybné a druhý ako bezchybné,
d) percento tých, ktoré prejdú oboma automatmi ako chybné.

1.6.9 Modifikujme situáciu z úlohy 1.6.8 takto: Predpokladajme, že druhý automat kontroluje iba tie výrobky, ktoré prvý automat označil ako chybné. Výstupná kontrola za nepodarky považuje tie výrobky, ktoré obidva automaty označia ako chybné. Predpokladajme, že (objektívna) nepodarkovosť výroby je 2 %-ná. Akú nepodarkovosť produkcii prisudzuje takáto výstupná kontrola?

1.7 Model nezávislého opakovania pokusu a Bernoulliho schéma

V tomto článku najprv vysvetlíme, ako modelujeme pokus, ktorý pozostáva z n-násobného nezávislého opakovania nejakého jednoduchého pokusu (odseky 1.7.1 až 1.7.5). Potom prejdeme k modelovaniu pokusu, ktorý pozostáva z dvoch (či viacerých) pokusov, vykonaných na sebe nezávisle. Obe situácie sú naozaj veľmi časté, a to nielen v pravdepodobnostných úlohách, ale aj v štatistike. Ako uvidíme, v štatistike majú kľúčovú úlohu.

1.7.1 Začneme konštrukciou modelu pre dvojnásobné opakovanie jednoduchého pokusu. Nech náš jednoduchý pokus má konečne veľa výsledkov (pre konkrétnosť štyri) a nech jeho model poznáme, nech je ním dvojica (Ω°, P°), kde

$$\Omega^\circ = (\omega_1^o,\ \omega_2^o,\ \omega_3^o,\ \omega_4^o)$$
$$P^\circ:\ (p_1^o,\ p_2^o,\ p_3^o,\ p_4^o).$$

Zápis $P^\circ : (p_1^o, p_2^o, p_3^o, p_4^o)$ hovorí, že P° je určená hodnotami $P^\circ(\omega_i^o)$, t. j. $p_i^o = P^\circ(\omega_i^o)$. Výsledky dvojnásobného (nezávislého) opakovania budú modelovať dvojice (ω_i^o, ω_j^o), čo znamená, že zložený pokus má 16 výsledkov:

$$\begin{array}{cccc} (\omega_1^o,\omega_1^o), & (\omega_1^o,\omega_2^o), & (\omega_1^o,\omega_3^o), & (\omega_1^o,\omega_4^o) \\ (\omega_2^o,\omega_1^o), & (\omega_2^o,\omega_2^o), & (\omega_2^o,\omega_3^o), & (\omega_2^o,\omega_4^o) \\ (\omega_3^o,\omega_1^o), & (\omega_3^o,\omega_2^o), & (\omega_3^o,\omega_3^o), & (\omega_3^o,\omega_4^o) \\ (\omega_4^o,\omega_1^o), & (\omega_4^o,\omega_2^o), & (\omega_4^o,\omega_3^o), & (\omega_4^o,\omega_4^o) \end{array}$$

Priestor výsledkov zloženého pokusu je teda množina $\Omega = \Omega^\circ \times \Omega^\circ$. Pravdepodobnosť v zloženom pokuse chceme definovať tak, aby tie udalosti, ktoré sú intuitívne nezávislé, boli nezávislé aj v zmysle definície nezávislosti z predchádzajúceho článku. Nie je ťažké postrehnúť, že pre pravdepodobnosť každého jednotlivého výsledku musí nutne platiť:

$$P((\omega_i^o, \omega_j^o)) = p_i^o p_j^o \tag{1}$$

To znamená, že uvedených 16 výsledkov zloženého pokusu bude mať pravdepodobnosti:

$$\begin{array}{cccc} p_1^o p_1^o, & p_1^o p_2^o, & p_1^o p_3^o, & p_1^o p_4^o \\ p_2^o p_1^o, & p_2^o p_2^o, & p_2^o p_3^o, & p_2^o p_4^o \\ p_3^o p_1^o, & p_3^o p_2^o, & p_3^o p_3^o, & p_3^o p_4^o \\ p_4^o p_1^o, & p_4^o p_2^o, & p_4^o p_3^o, & p_4^o p_4^o \end{array}$$

Pretože zložený pokus má konečne veľa výsledkov a každému z nich sme priradili pravdepodobnosť, je model zloženého pokusu stanovený. Ide o model *dvojnásobného nezávislého* opakovania (toho istého, jednoduchého) pokusu. Dá sa ľahko ukázať, že platí veta 1.7.2.

1.7.2 Veta. Nech (Ω°, P°) je model jednoduchého pokusu a P v zloženom pokuse je definovaná vzťahom (1). Ak A°, B° sú udalosti v jednoduchom pokuse a $P^\circ(A^\circ)$, $P^\circ(B^\circ)$ sú ich pravdepodobnosti, tak pre udalosti $A = A^\circ \times \Omega^\circ$, $B = \Omega^\circ \times B^\circ$, ktoré sú udalosťami v zloženom pokuse, platí:

$$P(A \cap B) = P(A^\circ \times B^\circ) = P^\circ(A^\circ)\,P^\circ(B^\circ) = P(A)\,P(B).$$

To znamená, že udalosti A, B, ktoré – intuitívne vzaté – sú iste nezávislé (veď A sa týka len prvého opakovania a B sa týka len druhého opakovania), sú nezávislé aj v zmysle formálnej definície nezávislosti z článku 1.6, veď nezávislosť znamená, že pre udalosti A, B platí: $P(A \cap B) = P(A)\,P(B)$. Všimnime si, že táto skutočnosť – nezávislosť udalostí $A^\circ \times \Omega^\circ$ a $\Omega^\circ \times B^\circ$ – vôbec nezávisí od toho, aké *konkrétne* hodnoty má pravdepodobnosť P° v bodoch (výsledkoch) $\omega_1^o, \omega_2^o, \omega_3^o, \omega_4^o$.

1.7.3 Príklad. Nech náhodný pokus spočíva v dvojnásobnom hádzaní obyčajnou hracou kockou. Zvoľme tri náhodné udalosti nášho zloženého pokusu takto:

A – v prvom hode menej ako 4 body,
B – v druhom hode aspoň 5 bodov,
C – v prvom hode menej ako 4 body, a pritom v druhom hode aspoň 5 bodov.

a) Stanovme pravdepodobnosti udalostí A, B, C. Zrejme $C = A \cap B$.

b) Intuitívne vzaté sú udalosti A, B nezávislé, pretože A sa týka prvého a B zase druhého hodu. Overme, že udalosti A, B sú nezávislé aj v zmysle formálnej definície, ak sme výsledkom zloženého pokusu priradili pravdepodobnosti podľa odseku 1.7.1.

Riešenie. Pretože kocka je riadna, podľa odseku 1.7.1 je pravdepodobnosť v zloženom pokuse stanovená vzťahom: $P(\omega_i^o, \omega_j^o) = p_i^o p_j^o = \frac{1}{36}$, pre $i, j = 1, 2, \dots, 6$, teda priestor zloženého pokusu je homogénny. Zrejme $|A| = 18$, $|B| = 12$, $|C| = |A \cap B| = 6$. Preto

$$P(A) = 18/36 = 1/2,\ P(B) = 12/36 = 1/3,\ P(C) = P(A \cap B) = 6/36 = 1/6.$$

b) Ako vidíme, skutočne platí $P(A \cap B) = P(A)P(B)$, teda udalosti A, B sú nezávislé.

1.7.4 Príklad. Náhodný pokus spočíva v dvojnásobnom hádzaní falošnou hracou kockou, ktorej steny (s bodmi 1 až 6) sa objavujú s pravdepodobnosťami: $\frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{2}{12}, \frac{3}{12}, \frac{4}{12}$. Zvoľme náhodné udalosti tak, ako v príklade 1.7.3, t. j.

A – v prvom hode menej ako 4 body,
B – v druhom hode aspoň 5 bodov,
C – v prvom hode menej ako 4 body, a pritom v druhom hode aspoň 5 bodov.

a) Stanovme pravdepodobnosti udalostí A, B a C.

b) Intuitívne sú zrejme udalosti A, B nezávislé, pretože A sa týka len prvého a B zase len druhého hodu. Overme, že udalosti A, B sú nezávislé aj v zmysle formálnej definície, ak sme výsledkom zloženého pokusu priradili pravdepodobnosti podľa odseku 1.7.1.

Riešenie. Pretože zložený pokus je opäť dvojnásobným nezávislým opakovaním hodu kockou, pravdepodobnosť v zloženom pokuse je opäť definovaná vzťahom $P(\omega_i^o, \omega_j^o) = p_i^o p_j^o$, pre $i, j = 1, 2, \dots, 6$. Tentoraz však priestor zloženého pokusu *nie je homogénny*, pretože kým výsledok $(1, 1)$ má pravdepodobnosť 1/144, tak výsledok $(6, 6)$ je 16-krát pravdepodobnejší! Ľahko sa zistí, že máme napr.

$$P(A) = P(\,(1,1), (1,2), \dots, (1,6), (2,1), (2,2), \dots, (2,6), (3,1), (3,2), \dots, (3,6)\,) =$$

$$= 3\left(\frac{1}{144} + \frac{1}{144} + \frac{1}{144} + \frac{2}{144} + \frac{3}{144} + \frac{4}{144}\right) = 3\,\frac{12}{144} = 3\,\frac{1}{12} = \frac{1}{4}.$$

Rovnako sa overí, že $P(B) = 7/12$ a $P(C) = 21/144 = 7/48$.

b) Udalosti A, B sú nezávislé, pretože platí $P(A \cap B) = P(C) = P(A)\,P(B)$. Naša intuícia je takto potvrdená. Nemáme dôvod pochybovať, že model pokusu je adekvátny.

1.7.5 Ak namiesto dvojnásobného opakovania jednoduchého pokusu uvažujeme trojnásobné, resp. viacnásobné opakovanie, prirodzeným spôsobom modifikujeme postup modelovania zloženého pokusu z odseku 1.7.1. Napríklad, pre $n = 3$, takto

výsledky zloženého pokusu sú usporiadané trojice $(\omega_i^o, \omega_j^o, \omega_k^o)$,
pre pravdepodobnosť týchto výsledkov kladieme $P(\omega_i^o, \omega_j^o, \omega_k^o) = p_i^o p_j^o p_k^o$

Úlohy 1.7.1, 1.7.2 predstavujú situácie, keď zložený pokus pozostáva z trojnásobného opakovania hodu kockou. Je užitočné, uvedomiť si rozdiel medzi úlohami 1.7.1 a 1.7.2.

1.7.6 Bernoulliho schéma. V zloženenom pokuse, ktorý pozostáva z n-násobného nezávislého opakovania jednoduchého pokusu nás často zaujímajú pravdepodobnosti len takých udalostí, ktoré vypovedajú o tom, *koľkokrát* nastala akási udalosť spojená s jednoduchým pokusom. Napríklad, uvažujme pokus, v ktorom hádžeme kockou päťkrát. Jednoduchým pokusom je takto hod kockou. V ňom zvoľme ako A, udalosť „na kocke padne aspoň 5 bodov“. Teraz v zloženom pokuse nás zaujíma udalosť spočívajúca v tom, že v sérii piatich hodov udalosť A nastane tri razy. Ak ide o bežnú hraciu kocku, tak pravdepodobnosť A sa rovná 1/3. Aká je pravdepodobnosť toho, že v piatich hodoch nastane A práve tri razy?

Pre úvahy v nasledujúcich kapitolách bude vhodné zaviesť označenie, ktoré nám pomôže už teraz. Spojme s i-tým opakovaním jednoduchého pokusu predstavu náhodnej veličiny X_i, ktorá je *identifikátorom* udalosti A v tom zmysle, že

ak v i-tom opakovaní A nastane, tak X_i má hodnotu 1 a
ak A v i-tom opakovaní nenastane, tak X_i má hodnotu 0.

Veličiny X_1, X_2, X_3, X_4, X_5 *identifikujú* udalosť nastatia, resp. nenastatia A v jednotlivých opakovaniach (jednoduchého) pokusu.

Ak počet hodov kockou $n = 5$ a ak A sa vyskytla práve v prvom, treťom a piatom opakovaní, tak $X_1 = X_3 = X_5 = 1$, kým $X_2 = X_4 = 0$. Formálne to môžeme zapísať takto:

$$(X_1, X_2, X_3, X_4, X_5) = (1, 0, 1, 0, 1).$$

Teraz však, ako sme povedali, zaujíma nás pravdepodobnosť udalosti spočívajúcej v tom, že sledovaná udalosť A nastane práve tri razy. To môžeme zapísať výrazom

$$X_1 + X_2 + X_3 + X_4 + X_5 = 3.$$

Našu udalosť, t. j. udalosť, že A nastane práve trikrát, realizuje týchto 10 výsledkov:

(1, 1, 1, 0, 0) (1, 1, 0, 1, 0) (1, 1, 0, 0, 1) (1, 0, 1, 1, 0) (1, 0, 1, 0, 1)
(1, 0, 0, 1, 1) (0, 1, 1, 1, 0) (0, 1, 1, 0, 1) (0, 1, 0, 1, 1) (0, 0, 1, 1, 1)

Pretože pravdepodobnosť každého z nich sa rovná $\left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^2$, tak z aditivity pravdepodobnosti máme

$$P(X_1 + X_2 + X_3 + X_4 + X_5 = 3) = 10 \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^2.$$

Vo všeobecnosti,

- ak sledovaná udalosť v jednoduchom pokuse má pravdepodobnosť p,
- ak ten jednoduchý pokus opakujeme nezávisle n-krát,
- ak ide o pravdepodobnosť udalosti v zloženom pokuse, ktorá spočíva v tom, že udalosť A nastane práve k-krát, t. j. ide o udalosť $\{X_1 + X_2 + \dots + X_n = k\}$, tak platí

$$P(X_1 + X_2 + \dots + X_n = k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{2}$$

Tento vzťah sa spája s pomenovaním *Bernoulliho schéma*, keďže vzorec sa týka takých pokusov, kde sú splnené tri uvedené predpoklady (t. j. pokusy majú rovnakú štruktúru).

1.7.7 Príklad. Normálnou mincou hodíme desaťkrát. Aká je pravdepodobnosť udalostí

a) A – na minci padne tri razy znak?
b) B – na minci padne sedem ráz znak?
c) C – na minci padne šesťkrát číslo?
d) D – minca ukáže postupne: (Z, Z, Z, Č, Č, Č, Z, Č, Č, Č)?
e) E – znak padne aspoň sedem ráz?

Riešenie. Sledovanou udalosťou je padnutie znaku, a tak $X_i = 1$ práve vtedy, keď v i-tom opakovaní padne znak. Zrejme $P(X_i = 1) = p = 0.5$ a pretože $n = 10$, máme

a) $P(A) = P(X_1 + X_2 + \dots + X_{10} = 3) = \binom{10}{3} 0.5^3 (1-0.5)^{10-3} = 120\,(0.5)^{10} = 0.1172$

b) $P(B) = P(X_1 + X_2 + \dots + X_{10} = 7) = \binom{10}{7} 0.5^7 (1-0.5)^{10-7} = 120\,(0.5)^{10} = 0.1172$

ako vidíme, vďaka tomu, že $p = 0.5$, $P(A) = P(B)$.

c) $P(C) = P(X_1 + X_2 + \dots + X_{10} = 4) = \binom{10}{4} 0.5^4 (1-0.5)^{10-4} = 210\,(0.5)^{10} = 0.2051$

d) Pozor, teraz ide o konkrétny jeden výsledok v zloženom pokuse

(Z, Z, Z, Č, Č, Č, Z, Č, Č, Č) = (1, 1, 1, 0, 0, 0, 1, 0, 0, 0), preto
$P(D) = p^4(1-p)^6 = 0.5^{10} = 1/1024 = 0.0009766$

e) $P(E) = \sum_{i=7}^{10} P(X_1 + X_2 + \dots + X_{10} = i) = 0.1172 + 0.0439 + 0.0098 + 0.0010 = 0.1719.$

1.7.8 Príklad. Vykonajme náhodný výber rozsahu $n = 100$ s vrátením. To znamená, že 100 ráz náhodne vytiahneme výrobok z produkcie, zistíme či je dobrý, alebo chybný a vrátime ho späť. Nech nepodarkovosť produkcie je $p100$ %-ná (to znamená, že náhodne vybratý výrobok je chybný s pravdepodobnosťou p). Určíme P(A), P(B), P(C) a P(D), keď

a) A – medzi 100 vybratými nie je chybný výrobok,
b) B – medzi 100 vybratými je práve jeden chybný výrobok,
c) C – medzi 100 vybratými sú práve tri chybné výrobky,
d) D – medzi 100 vybratými sú najviac tri chybné výrobky.

Riešenie. Sledovanou udalosťou je výskyt nepodarku a pravdepodobnosť tejto udalosti sa rovná p. Počet opakovaní pokusu $n = 100$.

a) Pretože ide o nezávislé opakovanie pokusu, $P(A) = (1-p)^{100}$.

b) $P(B) = 100p(1-p)^{99}$, pretože tým chybným môže byť hociktorý spomedzi 100 vybratých (to je 100 prípadov, ktoré sa navzájom vylučujú). Alebo, skrátka, podľa (2) odseku 1.7.6 ($k = 1$).

c) Tentoraz vo vzťahu (2), odseku 1.7.6 máme $k = 3$, preto $P(C) = \binom{100}{3} p^3(1-p)^{97}$

d) Teraz spočítame možnosti pre $k = 0, 1, 2, 3$, preto $P(D) = \sum_{i=0}^{3} \binom{100}{i} p^i(1-p)^{100-i}$

1.7.9 Doteraz zložený pokus spočíval v n-násobnom (nezávislom) opakovaní jednoduchého pokusu. Teraz sa venujme pokusu, ktorý je zložený z dvoch na sebe nezávislých pokusov. Výsledky prvého pokusu $x_1, x_2, \dots, x_m$ nech majú pravdepodobnosti $p_1, p_2, \dots, p_m$ a výsledky druhého označme $y_1, y_2, \dots, y_n$, ich pravdepodobnosti nech sú $q_1, q_2, \dots, q_n$.

Pre konkrétnosť, nech $m = 3$ a $n = 4$. Za výsledky zloženého pokusu berme usporiadané dvojice (x_i, y_j), a pretože pokusy sú nezávislé, máme mn možných výsledkov. Tak ako v odseku 1.7.1, model zloženého pokusu chceme vytvoriť tak, aby po formálnej stránke boli nezávislé také udalosti, ktoré sú intuitívne nezávislé, lebo prvá z nich sa týka len prvého pokusu a druhá len druhého pokusu. To dosiahneme definovaním pravdepodobnosti tak, ako ukazujú polia: vľavo je pole výsledkov zloženého pokusu a vpravo pole ich pravdepodobností:

$$\begin{matrix} (x_1,y_1) & (x_1,y_2) & (x_1,y_3) & (x_1,y_4) \\ (x_2,y_1) & (x_2,y_2) & (x_2,y_3) & (x_2,y_4) \\ (x_3,y_1) & (x_3,y_2) & (x_3,y_3) & (x_3,y_4) \end{matrix} \qquad \begin{matrix} p_1q_1 & p_1q_2 & p_1q_3 & p_1q_4 \\ p_2q_1 & p_2q_2 & p_2q_3 & p_2q_4 \\ p_3q_1 & p_3q_2 & p_3q_3 & p_3q_4 \end{matrix}$$

Pri takejto konštrukcii pravdepodobnosti zloženého pokusu platí veta 1.7.10.

1.7.10 Veta. Označme priestory výsledkov prvého a druhého pokusu symbolmi $\Omega^{(1)}$, resp. $\Omega^{(2)}$. Definujme P na súčine $\Omega^{(1)} \times \Omega^{(2)}$ tak, ako naznačuje schéma odseku 1.7.9, teda pre všetky i, j, kladieme

$$P((x_i, y_j)) = p_i q_j$$

Ak $A^\circ \subset \Omega^{(1)}$, $B^\circ \subset \Omega^{(2)}$, potom pre udalosti $A = A^\circ \times \Omega^{(2)}$ a $B = \Omega^{(1)} \times B^\circ$ (A, B sú udalosti v zloženom pokuse) platí:

$$P(A \cap B) = P(A)P(B).$$

(To znamená, že A, B sú nezávislé.)

Úlohy

1.7.1 Nech náhodný pokus spočíva v trojnásobnom hádzaní obyčajnou hracou kockou. Zvoľme náhodné udalosti

A – v prvom hode menej ako 4 body,
B – v druhom hode aspoň 5 bodov,
C – v treťom hode párny počet bodov.

a) Stanovte pravdepodobnosti udalostí A, B, C, $A \cap B$, $A \cap B \cap C$.
b) Overte, že udalosti A, B, C sú po dvoch nezávislé.
c) Sú udalosti A, B, C (ako trojica) nezávislé?

1.7.2 Nech náhodný pokus spočíva v trojnásobnom hádzaní falošnou hracou kockou, ktorej steny (s bodmi 1 až 6) sa objavujú s pravdepodobnosťami: $\frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{2}{12}, \frac{3}{12}, \frac{4}{12}$.
Zvoľme náhodné udalosti:

A – v prvom hode menej ako 4 body,
B – v druhom hode aspoň 5 bodov,
C – v treťom hode párny počet bodov.

a) Stanovte pravdepodobnosti udalostí A, B, C, $A \cap B$, $A \cap B \cap C$.
b) Overte, že udalosti A, B, C sú po dvoch nezávislé.
c) Sú udalosti A, B, C (ako trojica) nezávislé?

1.7.3 Falošnou mincou, na ktorej pravdepodobnosť padnutia znaku sa rovná 0.55, hodíme desaťkrát. Aká je pravdepodobnosť udalostí

a) A – na minci padne tri razy znak?
b) B – na minci padne sedemkrát znak?
c) C – na minci padne šesťkrát číslo?
d) D – minca ukáže postupne: (Z, Z, Z, Č, Č, Č, Z, Č, Č, Č)?
e) E – znak padne aspoň sedem ráz?

1.7.4 Falošnou kockou hádžeme päťkrát. Jednotlivé body (t. j. steny kocky) padajú s pravdepodobnosťami $\frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{2}{12}, \frac{3}{12}, \frac{4}{12}$. S akou pravdepodobnosťou

a) v nepárnych hodoch padne menej ako 4 a súčasne v párnych hodoch aspoň 5?
b) práve v troch hodoch padne menej bodov ako 4?
c) aspoň v troch hodoch padne menej bodov ako 4?

1.7.5 Náhodný pokus spočíva v šesťnásobnom hádzaní obyčajnou hracou kockou. S akou pravdepodobnosťou

a) nepadne šestka?
b) šestka padne práve dva razy?
c) šestka padne aspoň štyrikrát?

1.7.6 Uvažujme pokus z predchádzajúcej úlohy, ale teraz nech je kocka falošná a jej steny sa objavujú s pravdepodobnosťami $p_1, p_2, \dots, p_6$. Nech C znamená udalosť „padlo párne číslo". S akou pravdepodobnosťou

a) udalosť C nenastane ani raz v tých šiestich hodoch?
b) udalosť C nastane práve dva razy?
c) udalosť C nastane aspoň štyrikrát?

1.7.7 Predpokladajme, že pri výrobe tvarohových koláčikov sa dodržuje technológia, ktorá predpisuje použiť 10 000 hrozienok na 1000 koláčikov. Aká je pravdepodobnosť, že v zakúpenom koláčiku

a) nebude ani jedno hrozienko?
b) budú práve dve hrozienka?
c) bude menej ako 15, ale aspoň 7 hrozienok?

1.7.8 Predpokladajme, že dlhodobou evidenciou sa zistilo, že náhodné okolnosti s pravdepodobnosťou 0.12 znemožňujú študentovi prísť na prednášku. Ak celkový počet študentov sa rovná 100, s akou pravdepodobnosťou na náhodne vybranej prednáške

a) počet prítomných prekročil hodnotu 87?
b) počet prítomných nadobudol hodnotu z intervalu $\langle 70, 80)$?

1.7.9 Skúsenosť hovorí, že študent sa zúčastní na riadnom termíne skúšky s pravdepodobnosťou 0.92. Predpokladajme, že krúžok má 25 študentov. S akou pravdepodobnosťou bude skúšajúci skúšať

a) 23 študentov?
b) menej ako 23 študentov?

1.7.10 Skriptá majú 200 strán a na nich náhodne rozmiestnených 40 tlačových chýb (tým sa rozumie to, že konkrétna chyba sa môže dostať na ktorúkoľvek stranu s rovnakou pravdepodobnosťou). S akou pravdepodobnosťou na náhodne zvolenej strane

a) nie je tlačová chyba?
b) sú práve tri tlačové chyby?
c) sú najviac tri tlačové chyby?

1.7.11 Predpokladajme, že výrobky sa skúšajú pri preťažení, pričom každý z nich vydrží skúšku s pravdepodobnosťou 0.9. S akou pravdepodobnosťou zo série 100 výrobkov

a) skúšku nevydržalo práve 5 výrobkov?
b) počet tých, čo nevydržali skúšku, bol menší ako 10?

1.7.12 Zo štatistík je známe, že pravdepodobnosť narodenia chlapca sa rovná 0.515. S akou pravdepodobnosťou je z prvých 50 narodených detí nového roku aspoň 25 chlapcov?

1.7.13 Majme dve falošné kocky. Predpokladajme, že modrú kocku charakterizujú pravdepodobnosti $\frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{2}{12}, \frac{3}{12}, \frac{4}{12}$ a červenú pravdepodobnosti $\frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{2}{12}, \frac{2}{12}, \frac{5}{12}$. Náhodný pokus spočíva v hode týmito kockami. Aká je pravdepodobnosť toho, že

a) obe ukážu párne číslo?
b) na modrej bude párne a na červenej nepárne číslo?
c) na modrej aspoň 4 body a na červenej menej ako 5 bodov?

1.7.14 Na terč striedavo strieľajú dvaja strelci a zasahujú ho na sebe nezávisle s pravdepodobnosťami $p = 0.8$ a $q = 0.7$. Predpokladajme, že každý strieľa tri razy. Aká je pravdepodobnosť, že

a) prvýkrát je terč zasiahnutý druhou ranou prvého strelca?
b) terč bol zasiahnutý len raz a to druhou ranou prvého strelca?
c) terč bol zasiahnutý práve raz a to ranou prvého strelca?
d) terč bol zasiahnutý práve raz?
e) terč bol zasiahnutý práve dva razy?

1.7.15 Náhodný pokus spočíva v hode tromi falošnými mincami. Znaky padajú na jednotlivých minciach s pravdepodobnosťami 0.65, 0.55 a 0.45. S akou pravdepodobnosťou padne znak práve na dvoch minciach?

2 NÁHODNÉ VELIČINY

2.1 Diskrétne náhodné veličiny a ich rozdelenia

Náhodná veličina predstavuje číselný popis výsledku náhodného pokusu. V pokuse s dvoma hracími kockami je náhodnou veličinou napr. súčet padnutých bodov. Ďalšími veličinami (stále v rámci pokusu s dvoma kockami) sú napr. súčin, alebo rozdiel padnutých bodov. Budeme ich označovať veľkými písmenami (obyčajne z konca abecedy), takže môžeme hovoriť o veličinách X, Y, resp. Z. Po matematickej stránke je náhodná veličina zobrazenie. Ak (Ω, S, P) je pravdepodobnostný priestor, tak náhodná veličina X je zobrazenie z Ω do R, čo zapisujeme známym spôsobom

$$X: \Omega \to \mathrm{R}.$$

Aj keď X je zobrazenie, budeme s ním pracovať trochu inak ako napr. v matematickej analýze. Budú nás zaujímať možné hodnoty veličiny X a tiež pravdepodobnosti, s akými veličina X svoje hodnoty nadobúda. Ak množina možných hodnôt veličiny X je konečná (prípadne nekonečná, ale spočítateľná), hovoríme, že X je *diskrétna* náhodná veličina. V takom prípade nás zaujíma *rozdelenie pravdepodobnosti* (krátko: *rozdelenie*), ktoré intuitívne chápeme ako rozloženie jednotkovej pravdepodobnosti na všetky potenciálne možné hodnoty veličiny X. Je dôležité uvedomovať si rozdiel medzi veličinou X ako takou a jej možnými hodnotami – *realizáciami*, ktoré budeme označovať x_i. Množinu všetkých potenciálne možných realizácií veličiny X označujeme $H(X)$, $H(X) = \{x_1, x_2, x_3, \dots\}$.

Ak X je napr. súčet bodov na dvoch riadnych kockách, zrejme $H(X) = \{2, 3, \dots, 11, 12\}$. Rozdelenie veličiny X predstavujú hodnoty $P(\{X = 2\})$, $P(\{X = 3\})$, ... , $P(\{X = 12\})$. Intuitívne je zrejmé, že sú to pravdepodobnosti náhodných udalostí. Napr. udalosť $\{X = 2\}$ nastane práve vtedy, keď na oboch kockách padne jeden bod. To je jeden výsledok z 36 rovnako pravdepodobných výsledkov. Preto $\mathrm{P}(\{X = 2\}) = 1/36$. V ďalšom namiesto $\mathrm{P}(\{X = x\})$ budeme písať jednoducho $\mathrm{P}(X = x)$.

Pre stanovenie $\mathrm{P}(X = 7)$ pomôže, keď X vidíme ako zobrazenie, ktoré pracuje takto: Ak výsledok hodu kockami označíme (i, j), t. j. $\omega = (i, j)$, tak $X(\omega) = X((i, j)) = i + j$. Potom napr. pre pravdepodobnosť udalosti $\{X = 7\}$ platí

$$\mathrm{P}(X = 7) = \mathrm{P}(\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}) = \tfrac{6}{36}.$$

Jednotková pravdepodobnosť sa rozdelí na 11 možných hodnôt veličiny X a je pohodlné toto rozdelenie pravdepodobnosti prezentovať v tvare tabuľky:

$H(X)$	2	3	4	5	6	7	8	9	10	11	12
$\mathrm{P}(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Tabuľka predstavuje *rozdelenie pravdepodobnosti* veličiny X, krátko: *rozdelenie* veličiny X.

V definícii náhodnej veličiny, ktorá nasleduje, budeme hovoriť o náhodných udalostiach spočívajúcich v tom, že veličina X sa realizuje hodnotou, ktorá leží v intervale $\langle a, b)$. Takú udalosť zapisujeme ako $\{a \le X < b\}$ a jej pravdepodobnosť ako $P(\{a \le X < b\})$. Opäť budeme trochu nedôslední a namiesto $P(\{a \le X < b\})$ píšeme jednoducho $P(a \le X < b)$. Dôležité je uvedomiť si, že $\{a \le X < b\}$ je úsporný zápis pre $\{\omega: a \le X(\omega) < b\}$, teda pre udalosť, podmnožinu priestoru Ω. V našej situácii, keď X predstavuje súčet bodov na kockách, náhodnú udalosť $\{5 \le X < 8\}$ vytvára 15 prvkov priestoru Ω (štyri z nich X zobrazuje do hodnoty 5, ďalších päť do hodnoty 6 a šesť bodov priestoru Ω veličina X zobrazuje do hodnoty 7).

2.1.1 Definícia. Nech (Ω, S, P) je pravdepodobnostný priestor a X je zobrazenie z Ω do R. Zobrazenie X nazývame *náhodná veličina*, ak má zmysel hovoriť o $P(a \le X < b)$, pre akékoľvek $a, b \in R$, $a < b$. Hovoríme, že X je *diskrétna* náhodná veličina, ak nadobúda konečne, alebo spočítateľne veľa hodnôt x_i, pričom každú z nich nadobúda s kladnou pravdepodobnosťou. Funkciu f

$$f: R \to \langle 0, 1\rangle$$

definovanú vzťahom

$$f(x) = P(X = x)$$

nazývame *pravdepodobnostná funkcia* veličiny X. Hodnoty x_i, pre ktoré platí $f(x_i) > 0$ nazývame *možné hodnoty* (tiež *realizácie*) veličiny X a množinu všetkých možných hodnôt X označujeme $H(X)$.

2.1.2 Príklad. V náhodnom pokuse, ktorý spočíva v hode dvoma obyčajnými hracími kockami, nech Y znamená absolútnu hodnotu rozdielu bodov, ktoré padli na kockách. Určte pravdepodobnostnú funkciu veličiny Y. V rámci tohoto pokusu uvažujme aj veličiny U, V, pričom U predstavuje minimum z bodov, ktoré padli na kockách a V predstavuje maximum z bodov, ktoré padli na kockách. Určte pravdepodobnostné funkcie veličín U, V. Všimnime si, že ak X predstavuje súčet bodov, ktoré padli na kockách, tak platí: $X = U + V$.

2.1.3 Veta. Nech X je diskrétna náhodná veličina, $H(X) = \{x_1, x_2, \dots, x_n, \dots\}$ je množina jej možných hodnôt a f je pravdepodobnostná funkcia veličiny X. Potom platí:

1. $f(x) \ge 0$ pre všetky $x \in R$ a $\sum_{x_i \in H(X)} f(x_i) = 1$,
2. $P(a \le X < b) = \sum_{a \le x_i < b} f(x_i)$,
3. $P(X < c) = \sum_{x_i < c} f(x_i)$.

2.1.4 Poznámka. Niekedy sa pracuje s pojmom pravdepodobnostná funkcia bez toho, aby bol definovaný pravdepodobnostný priestor. V takom prípade sa pravdepodobnostná funkcia definuje ako funkcia

$$f: R \to \langle 0, 1\rangle$$

nenulová len v bodoch konečnej (prípadne spočítateľnej) množiny H, pre ktorú platí:

$$\sum_{x \in H} f(x) = 1.$$

Dá sa ukázať, že ak f má uvedené vlastnosti, tak je možné formálne definovať priestor (Ω, S, P) a definovať na ňom X tak, aby f bola pravdepodobnostnou funkciou zostrojenej veličiny X. Preto často budeme zastávať postoj, že rozdelením, teda v tomto prípade

pravdepodobnostnou funkciou, je diskrétna náhodná veličina definovaná. Takýto prístup ilustruje nasledujúci príklad.

2.1.5 Príklad. Nech $n \in N$ je prirodzené číslo a funkcia f nech je daná vzťahom:

$$f(i) = \begin{cases} k2^i, & i = 0, 1, \dots, n-1, \\ 0, & inak. \end{cases}$$

Určme konštantu k tak, aby f mohla byť považovaná za pravdepodobnostnú funkciu nejakej náhodnej veličiny.

Riešenie. Pretože má platiť $f(x) \ge 0$ pre všetky $x \in R$, musí byť $k \ge 0$. Zrejme prípad $k = 0$ vylúčime, veď identicky nulová funkcia nemôže byť pravdepodobnostnou funkciou. Preto nutne $k > 0$. Ďalšie ohraničenie pre k získame z podmienky, že súčet všetkých hodnôt $f(x_i)$ sa musí rovnať 1. Pretože

$$\textstyle\sum_{i=0}^{n-1} f(x_i) = k(2^0 + 2^1 + 2^2 + \dots + 2^{n-1}) = k(2^n - 1),$$

z podmienky $k(2^n - 1) = 1$ máme: $k = \frac{1}{2^n - 1}$. To znamená, že bude platiť: $P(X = i) = \frac{2^i}{2^n - 1}$ a napr. pre $n = 5$, tabuľka f má tvar

x_i	0	1	2	3	4
$f(x_i)$	$\frac{1}{31}$	$\frac{2}{31}$	$\frac{4}{31}$	$\frac{8}{31}$	$\frac{16}{31}$

2.1.6 Príklad. Nech funkcia f je daná vzťahom:

$$f(k) = \begin{cases} \dfrac{1}{k(k+1)}, & k = 1, 2, \dots, n, \dots \\ 0, & inak. \end{cases}$$

Ukážme, že f má vlastnosti pravdepodobnostnej funkcie, a preto má zmysel uvažovať o náhodnej veličine X, pre ktorú bude f pravdepodobnostnou funkciou. Ako vidíme, množina možných hodnôt veličiny X je nekonečná, $H(X) = \{1, 2, \dots, n, \dots\}$. Určme $P(X < 4)$.

Riešenie. Funkciu môžeme zapísať pohodlne tabuľkou, ktorá uvádza hodnoty v tých argumentoch, v ktorých je f kladná (v ostatných argumentoch sa f rovná nule).

k	1	2	3	...	n	...
$f(k)$	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{12}$	...	$\frac{1}{n(n+1)}$	...

Počítajme súčet hodnôt $f(k)$ pre $k = 1, 2, \dots, n, \dots$ a presvedčme sa, že sa rovná 1:

$$\frac{1}{2} + \frac{1}{6} + \frac{1}{12} + \dots + \frac{1}{n(n+1)} + \dots = \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \dots + \left(\frac{1}{n} - \frac{1}{n+1}\right) + \dots = 1.$$

Nakoniec, $P(X < 4) = P(X = 1) + P(X = 2) + P(X = 3) = \frac{1}{2} + \frac{1}{6} + \frac{1}{12} = \frac{3}{4}$

V prvých príkladoch tohoto článku sme rozdelenie X určili deduktívnou cestou, keď veličina X predstavovala súčet bodov na kockách, alebo maximum, či minimum bodov na kockách). V príkladoch 2.1.5, 2.1.6 bolo rozdelenie viac-menej umelo dané pravdepodobnostnou funkciou. Teraz uvažujme o celkom reálnej situácii.

2.1.7 Príklad. Automat zvára blatník auta na piatich miestach. Kontrolou 100 blatníkov sme zistili, že na 79 sa nenašiel ani jeden nekvalitný zvar, jeden nekvalitný zvar bol zistený na 14 blatníkoch, dva nekvalitné zvary na štyroch blatníkoch, tri na dvoch a štyri nekvalitné zvary v jednom prípade. Nech X predstavuje počet nekvalitných zvarov. V takomto prípade aproximujeme rozdelenie veličiny X tabuľkou:

$H(X)$	0	1	2	3	4	5
Odhady $P(X = i)$	0.79	0.14	0.04	0.02	0.01	0

Niekoľko ďalších príkladov diskrétnych veličín z bežného života: počet ľudí čakajúcich na obsluhu v okamihu, keď sa zaradíme medzi čakajúcich (na pošte, alebo v banke), alebo počet áut, ktoré prejdú okolo nášho stanovišťa (ak sme účastníci prieskumu, ktorý mapuje dopravnú situáciu) a pod. Dôležitú úlohu hrajú diskrétne veličiny aj v popise spoľahlivosti výrobných procesov, pretože napr. počet nepodarkov vyrobených počas pracovného dňa je zaiste náhodný. V takýchto reálnych situáciách rozdelenie náhodnej veličiny nepoznáme a práve pozorovaním sa snažíme získať o ňom predstavu. Na základe pozorovaní aproximujeme pravdepodobnosť $P(X = k)$ *pomernou početnosťou* udalosti $\{X = k\}$ tak, ako sme to urobili v tomto príklade.

Úlohy

2.1.1 Nech pokus spočíva v hode štyrmi riadnymi mincami a veličina X predstavuje počet mincí, na ktorých padol znak. Pre $H(X)$ máme $H(X) = \{0, 1, 2, 3, 4\}$. Nájdite rozdelenie X, teda pravdepodobnostnú funkciu f veličiny X.

2.1.2 Pokus spočíva v hode šiestimi normálnymi hracími kockami a nech veličina X predstavuje počet tých kociek, na ktorých padlo aspoň 5 bodov. Teraz $H(X) = \{0, 1, 2, \ldots, 6\}$. Nájdite pravdepodobnostnú funkciu f veličiny X. Určte $P(X < 4)$.

2.1.3 Predpokladajme, že nepodarkovosť výroby je 2 %-ná. Nech veličina X predstavuje počet nepodarkov v sérii 50 výrobkov. Nájdite pravdepodobnostnú funkciu f veličiny X. Určte $P(X > 3)$.

2.1.4 V škatuli máme 10 výrobkov a medzi nimi sú tri nepodarky. Vytiahnime náhodne štyri a to:

a) naraz štyri
b) postupne jeden za druhým, bez vrátenia vybratých
c) postupne jeden za druhým, s vrátením vybratého (pred ťahaním ďalšieho)

Nech X, Y, Z sú veličiny predstavujúce počet chybných medzi vybratými (vo verzii (a), (b), resp. (c)). Určte rozdelenie veličín X, Y, Z, t. j. nájdite ich pravdepodobnostné funkcie.

2.1.5 V škatuli je 10 lístkov, očíslovaných od 1 do 10. Náhodne vyberáme naraz dva lístky. Nech X je veličina, ktorá predstavuje absolútnu hodnotu rozdielu čísel na vytiahnutých lístkoch. Nájdite rozdelenie X a určte $P(2 \le X < 6)$.

2.1.6 Pokus spočíva v hádzaní piatimi mincami, ktorých ruby, resp. líca sme označili číslami jedna, resp. dva. Nech X predstavuje súčet čísel, ktoré padli na minciach. Nájdite tabuľku rozdelenia veličiny X a určte $P(3 \le X < 8)$.

2.1.7 Na terč striedavo strieľajú dvaja strelci a zasahujú ho na sebe nezávisle s pravdepodobnosťami $p_1 = 0.8$ a $p_2 = 0.7$. Predpokladajme, že každý strieľa 2 razy. Nech X je celkový počet zásahov. Nájdite rozdelenie veličiny X a určte $P(X \ge 3)$.

2.1.8 Predpokladajme, že tri súčiastky zlyhávajú počas doby T s danými pravdepodobnosťami $p_1 = 0.3$, $p_2 = 0.4$ a $p_3 = 0.5$, pričom udalosti zlyhania sú nezávislé. Nech X označuje počet zlyhaných počas doby T. Určte rozdelenie veličiny X.

2.1.9 V škatuli máme 10 výrobkov a medzi nimi sú štyri nepodarky. Náhodne vyťahujeme výrobok, až kým nevytiahneme dobrý (vytiahnuté výrobky späť nevraciame). Nech X je počet vytiahnutých v pokuse. Určte rozdelenie veličiny X.

2.1.10 Uvažujme o obmene predchádzajúcej úlohy – pokračujeme vo výbere, až kým nevytiahneme dva dobré. Nech X znamená počet vytiahnutých výrobkov v pokuse. Určte rozdelenie veličiny X.

2.1.11 V škatuli máme päť loptičiek s číslami 1, 2, … , 5. Náhodne vyberieme naraz tri. Nech X je najväčšie z vytiahnutých čísel, Y je najmenšie a Z nech je súčet čísel na vytiahnutých loptičkách. Nájdite rozdelenie veličín X, Y a Z.

2.1.12 Z balíčka 52 francúzskych kariet sme náhodne vybrali 5 kariet (výberom bez vrátenia). Nech X predstavuje počet figúr medzi vybranými (keďže za figúry považujeme J, Q, K, A, balíček obsahuje 16 figúr). Nájdite rozdelenie veličiny X.

2.1.13 Hráč hrá hru, v ktorej s pravdepodobnosťou p získava bod a s pravdepodobnosťou q bod stráca (teda získava -1 bod), pričom zrejme $p + q = 1$. Nech X predstavuje hráčovo skóre po piatich kolách tejto hry. Nájdite rozdelenie veličiny X. (Ak napr. hráč bol dva razy úspešný a tri razy neúspešný, $X = -1$).

2.1.14 Milan a Cyril hrajú hazardnú hru: Milan hádže modrou a Cyril červenou kockou. 10 centov získava od súpera ten, komu padne väčšie číslo (ak padnú rovnaké čísla, hra končí remízou a nikto nikomu nič neplatí). Predpokladajme, že každý z nich začína hrať s 30-timi centami (teda začiatočný kapitál každého je 30 centov). Nech X predstavuje kapitál Milana po troch kolách tejto hry. Nájdite rozdelenie veličiny X.

2.1.15 Nech náhodný pokus spočíva v 6-násobnom nezávislom opakovaní hodu riadnou mincou. Za výsledok pokusu považujeme usporiadanú 6-ticu symbolov C, Z (číslo, znak). Keďže minca je riadna, t. j. symetrická a homogénna, každý zo 64 výsledkov má rovnakú pravdepodobnosť. Nech X predstavuje počet *zmien* v 6-tici. Napr. ak výsledok pokusu je ω, ω = (C, Z, Z, C, Z, C), tak $X(\omega) = 4$. Keby výsledkom bol (C, Z, Z, Z, C, C), tak X sa realizuje hodnotou 2. Zrejme X je náhodná veličina a $H(X) = \{0, 1, \ldots, 5\}$. Určte rozdelenie veličiny X. Viete nájsť rozdelenie X aj pre prípad n-násobného hodu (riadnou) mincou?

2.1.16 (Úloha je náročnejšia, ale zaujímavá.) Nech X predstavuje počet *zhôd* v náhodnej permutácii cifier $1, 2, \dots, 9$ (každá možná permutácia je rovnako pravdepodobná). Napr. ak výsledkom náhodného permutovania je permutácia (2, 5, 1, 8, 7, 9, 3, 4, 6), tak X sa realizuje hodnotou 0. Keby výsledkom bola permutácia (2, 5, 1, 4, 7, 9, 3, 8, 6), tak X sa realizuje hodnotou 2, pretože dve cifry – 4 a 8 – sú na svojich miestach. Zrejme $H(X)$ sa rovná $\{0, 1, 2, \dots, 6, 7, 9\}$, lebo $8 \notin H(X)$. Nájdite rozdelenie náhodnej veličiny X.

2.1.17 Uvažujme o náhodnom pokuse, ktorý pozostáva z opakovaného hádzania mincou dovtedy, kým sa nedočkáme padnutia znaku. Nech náhodná veličina X predstavuje počet hodov v pokuse. Zrejme $H(X) = \{1, 2, 3, \dots\}$. Nájdite rozdelenie veličiny X, teda jej pravdepodobnostnú funkciu.

2.1.18 Uvažujme o obmene predchádzajúcej úlohy, keď teraz je minca falošná a taká, že padnutie znaku má pravdepodobnosť 0.6. Aké je v tomto prípade rozdelenie veličiny X?

2.1.19 Uvažujme o náhodnom pokuse, ktorý pozostáva z opakovaného hodu normálnou hracou kockou dovtedy, kým sa nedočkáme padnutia čísla väčšieho ako 4 (t. j. kým sa nedočkáme aspoň piatich bodov). Nech veličina X predstavuje počet hodov v pokuse. Nájdite pravdepodobnostnú funkciu veličiny X.

2.1.20 Hádžeme normálnou hracou kockou dovtedy, až kým kocka neukáže dva razy 6 bodov (nie nutne za sebou). Nech X je veličina, ktorá predstavuje počet hodov v pokuse. Nájdite jej pravdepodobnostnú funkciu.

2.2 Spojité náhodné veličiny a ich rozdelenia

Doteraz sme hovorili o náhodných veličinách, ktorých množina možných hodnôt bola konečná, alebo ak nekonečná, tak spočítateľná. Ako vieme, také veličiny nazývame *diskrétne*. Na druhej strane sa často stretávame so situáciami, keď náhodným je napr. čas, ktorý trávime čakaním na obsluhu (čakaním vo fronte, či už v supermarkete, v banke, alebo na pošte), resp. náhodným je čas bezporuchového fungovania hociakého výrobku. V týchto prípadoch ide o náhodné veličiny, ktorých potenciálne hodnoty sú akékoľvek body intervalu, či už ohraničeného, alebo neohraničeného (často berieme interval $(0, \infty)$ len preto, že potom sa s modelom dobre pracuje). Takéto náhodné veličiny budeme nazývať *spojité* (presná definícia nasleduje).

Pripomeňme známy fakt – každý interval kladnej dĺžky je nespočítateľná množina. Nuž a to je dôvod, pre ktorý pravdepodobnostný opis spojitej náhodnej veličiny nie je možné urobiť pomocou pravdepodobností jednotlivých výsledkov (tých možných výsledkov je totiž nespočítateľne veľa, voľne povedané, príliš veľa) a naviac – a to vyzerá ako paradox – pre pravdepodobnosti jednotlivých výsledkov platí

$$P(X = c) = 0, \text{ pre každé reálne } c.$$

2.2.1 Definícia. Nech (Ω, S, P) je pravdepodobnostný priestor a X je zobrazenie z Ω do R. Zobrazenie X nazývame náhodná veličina, ak pre všetky $a, b \in R$, $a < b$, má zmysel hovoriť o $P(\{\omega: a \le X(\omega) < b\})$, to znamená, že množina $\{\omega: a \le X(\omega) < b\}$ patrí do S.

2.2.2 Poznámka. To, že $\{\omega: a \le X(\omega) < b\}$ je náhodná udalosť, t. j. $\{\omega: a \le X(\omega) < b\} \in S$, budeme brať ako samozrejmosť a nebudeme to overovať. Podstatné je, že rozumieme zápisu $P(\{\omega: a \le X(\omega) < b\})$ a chápeme ho takto:

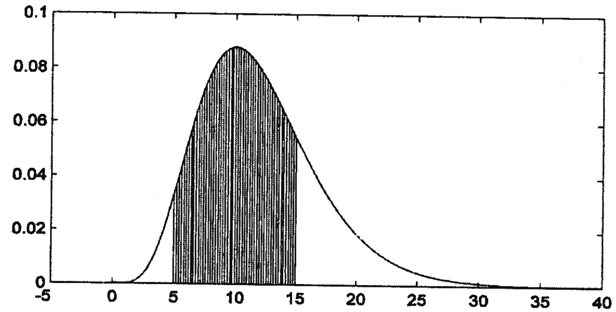
> $P(\{\omega: a \le X(\omega) < b\})$ je pravdepodobnosť javu, ktorý spočíva v tom, že v pokuse sa X realizuje hodnotou $x = X(\omega)$, ktorá leží v intervale $\langle a, b)$, t. j. $a \le X(\omega) < b$.

Pre $P(\{\omega: a \le X(\omega) < b\})$ budeme používať jednoduchší zápis: $P(a \le X < b)$. Ak pre všetky $a, b \in R$ vieme stanoviť $P(a \le X < b)$, hovoríme, že poznáme *rozdelenie* veličiny X.

2.2.3 Definícia. Nech (Ω, S, P) je pravdepodobnostný priestor a X je náhodná veličina. Hovoríme, že X je spojitá veličina s hustotou, ak pravdepodobnosti $P(a \le X < b)$ sú dané prostredníctvom nezápornej, po častiach spojitej funkcie f, $f: R \to \langle 0, \infty)$ tak, že platí

$$P(a \le X < b) = \int_a^b f(x)dx \quad \text{pre všetky } a, b \in R,\ a < b.$$

Funkcia f sa nazýva *hustota* veličiny X, resp. hustota rozdelenia, ktoré má veličina X.



Obr. 2-1. $P(5 \le X < 15) = \int_5^{15} f(x)\,dx$

2.2.4 Veta.

1. Ak f je hustota X, tak platí: $f(x) \ge 0$, pre všetky $x \in R$ a $\int_{-\infty}^{\infty} f(x)\,dx = 1$.
2. Ak X je veličina s hustotou, tak platí: $P(X = c) = 0$, pre každé $c \in R$.
3. Ak $f: R \to \langle 0, \infty)$ je po častiach spojitá a platí $\int_{-\infty}^{\infty} f(x)\,dx = 1$, tak je možné zostrojiť priestor (Ω, S, P) a na ňom definovať veličinu X tak, že f bude hustotou X. (To je dôvod, prečo o hustote môžeme hovoriť aj vtedy, keď žiaden (Ω, S, P) ani veličina X nie sú naporúdzi.)

2.2.5 Poznámka. Z bodu 2 vety 2.2.4 plynie dôležitá skutočnosť: Ak X je veličina s hustotou, tak

$$P(a \le X < b) = P(X = a) + P(a < X < b) = P(a < X < b) = P(a < X \le b) = P(a \le X \le b).$$

Vo všetkých príkladoch a úlohách bude hustota f daná ako funkcia, ktorá je nenulová (teda kladná) na ohraničenom, alebo aj neohraničenom intervale (prípadne kladná na celej číselnej osi). Bude praktické označiť ten interval symbolom $H(X)$ a chápať ho ako množinu možných hodnôt veličiny X, teda $H(X) = \{x \in R: f(x) > 0\}$.

2.2.6 Príklad. Určime hodnotu konštanty a tak, aby funkcia f definovaná vzťahom

$$f(x) = \begin{cases} a + x, & 0 < x < 1, \\ 0, & inak, \end{cases}$$

bola hustota. Zrejme nutne musí byť $a = \frac{1}{2}$, čo dostávame z podmienky $\int_{-\infty}^{\infty} f(x)dx = 1$. Keby však f bola definovaná vzťahom

$$f(x) = \begin{cases} a + x, & 1 < x < 2, \\ 0, & inak, \end{cases}$$

tak musí platiť $a = -\frac{1}{2}$. V prvom prípade $H(X) = (0, 1)$, v druhom $H(X) = (1, 2)$.

2.2.7 Príklad. Rozhodnime o úlohe symbolu a vo funkcii f definovanej vzťahom

$$f(x) = \begin{cases} \dfrac{2x}{a^2}, & 0 < x < a, \\ 0, & inak. \end{cases}$$

Určime $P(X < \frac{a}{2})$.

Riešenie. Funkcia f je po častiach spojitá (načrtnite jej graf) a zrejme $f(x) \geq 0$ pre všetky $x \in \mathrm{R}$. Teraz ide o to, pre akú hodnotu a sa integrál z funkcie f bude rovnať 1. Počítajme

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_0^a \frac{2x}{a^2}\,dx = \frac{2}{a^2}\int_0^a x\,dx = \frac{2}{a^2}\left[\frac{x^2}{2}\right]_0^a = \frac{a^2}{a^2} = 1.$$

Ako vidíme, integrál sa rovná jednej pre ľubovoľné kladné a. Hovoríme, že symbol a má úlohu parametra v uvažovanom rozdelení. Parameter a môže mať akúkoľvek kladnú hodnotu. Nakoniec

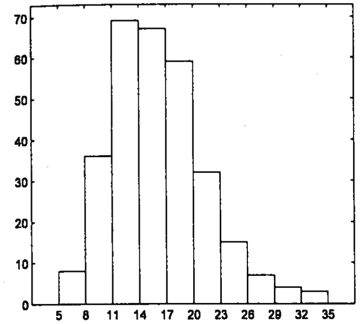
$$P\left(X < \frac{a}{2}\right) = \int_{-\infty}^{\frac{a}{2}} f(x)dx = \int_{-\infty}^{0} 0\,dx + \int_0^{\frac{a}{2}} \frac{2x}{a^2}\,dx = 0 + \frac{1}{a^2}\,[x^2]_0^{\frac{a}{2}} = \frac{1}{4}.$$

Je prirodzené pýtať sa, kedy máme do činenia s rozdeleniami, ktoré obsahujú parametre. Snáď ešte prirodzenejšia je otázka: Odkiaľ sa vôbec hustoty berú? Na základe čoho vieme, že rozdelenie skúmanej veličiny je určené konkrétnou hustotou?

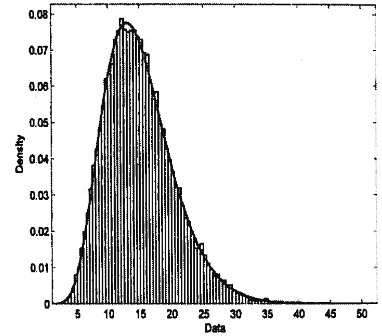
Odpoveď na prvú otázku dostaneme v článku 2.4, kde hovoríme o špeciálnych rozdeleniach. Teraz sa venujme druhej otázke: Na základe čoho v reálnej situácii skúmané rozdelenie modelujeme konkrétnou hustotou?

2.2.8 Príklad. Uvažujme o veličine X, ktorá predstavuje čas, potrebný na obsluhu zákazníka napr. v banke. Predstavme si, že pozorujeme niekoľko prepážok vo veľkej banke, ktoré vybavujú bežných klientov. Počas pracovného dňa sa tam realizujú stovky operácií a keby sme mali dobre organizované pozorovanie, na konci pracovnej doby by sme dostali konkrétne dáta, ktoré predstavujú (náhodné) časy obsluhy (časy potrebné na založenie účtu, na vklad, resp. výber z účtu a pod.). Predpokladajme, že behom pracovného dňa sme získali 300 realizácii náhodnej veličiny X. Tých 300 realizácií sme zobrazili histogramom, ktorý pracoval s desiatimi triedami (pozri obr. 2-2a). Početnosti tried vieme čítať na osi y, sú nimi čísla: 8, 36, 69, 67, 59, 32, 15, 7, 4, 3.

Obrázok 2-2b odpovedá situácii, v ktorej je histogram zostrojený z 3000 dát, a preto triedy môžu byť užšie a je ich viac (70). Vďaka tomu histogram vykresľuje tvar rozdelenia oveľa vernejšie. Hustotu intuitívne chápeme ako limitný prípad, ktorý by sme dostali, keby počet dát mohol konvergovať do nekonečna a súčasne by šírky tried konvergovali k nule (na obrázku je mierka na osi y zvolená tak, aby bolo možné znázorniť aj funkciu hustoty).



Obr. 2-2a. Histogram z 300 realizácií



Obr. 2-2b. Histogram z 3000 realizácií

2.2.9 Poznámka. Na záver poznamenajme, že funkcia hustoty nie je jednoznačná. Podľa definície je totiž hustotou každá taká nezáporná (po častiach spojitá) f, že platí

$$P(a \leq X < b) = \int_a^b f(x)dx \quad \text{pre všetky } a, b \in \mathrm{R},\ a < b.$$

Ak funkciu g získame z funkcie f tak, že f zmeníme len v konečne mnohých argumentoch, tak aj g bude hustotou X. To preto, lebo z matematickej analýzy je známe, že ak f a g sa líšia len v konečne veľa bodoch, tak platí

$$\int_a^b f(x)dx = \int_a^b g(x)dx \quad \text{pre všetky } a, b \in \mathrm{R},\ a < b.$$

Ak teda $g(x) \geq 0$, pre všetky $x \in \mathrm{R}$, aj funkcia g je hustota X. Nejednoznačnosť sa odohráva obyčajne v bodoch nespojitosti, ktorými sú okraje intervalu, ktorý predstavuje množinu možných hodnôt $H(X)$. Vezmime na vedomie, že nie je podstatné, ako definujeme hustotu v krajných bodoch intervalu $H(X)$.

Úlohy

2.2.1 Nech funkcia hustoty f je daná vzťahom

a) $f(x) = \begin{cases} kx, & 0 < x < 1, \\ 0, & inak. \end{cases}$ b) $f(x) = \begin{cases} kx, & 0 < x < 2, \\ 0, & inak. \end{cases}$

Rozhodnite, či k je konkrétna konštanta (v takom prípade určte jej hodnotu), alebo parameter (v takom prípade určte množinu možných hodnôt parametra).

2.2.2 Nech funkcia hustoty f je daná vzťahom

$$f(x) = \begin{cases} ax^{a-1}, & 0 \leq x \leq 1, \\ 0, & inak. \end{cases}$$

Rozhodnite o úlohe symbolu a. Je to konkrétna hodnota alebo parameter? V prvom prípade určite tú konkrétnu hodnotu. Ak usúdite, že a môže byť parameter, určite množinu prípustných hodnôt parametra a.

2.2.3 Nech funkcia hustoty f je daná vzťahom

$$f(x) = \begin{cases} a(1-|x|), & -1 < x < 1, \\ 0, & inak. \end{cases}$$

a) Rozhodnite o úlohe symbolu a. Je to parameter, alebo konkrétna hodnota?
b) Nájdite $P(|X| < 0.5)$, t. j. určte $P(-0.5 < X < 0.5)$.
c) Určte m tak, aby $P(X < m) = 0.5$.
d) Určte c tak, aby $P(X < c) = 0.9$.

2.2.4 Nech funkcia hustoty f je daná vzťahom

$$f(x) = \begin{cases} ax, & 0 < x \le 1, \\ a, & 1 < x < 2, \\ 0, & inde. \end{cases}$$

a) Rozhodnite o úlohe symbolu a. Je to parameter, alebo konkrétna hodnota?
b) Nájdite $P(0.5 < X < 1.5)$.
c) Určte c tak, aby $P(X < c) = 0.9$.

2.2.5 Nech funkcia hustoty f je daná vzťahom

$$f(x) = \begin{cases} ax^2, & 0 < x < 1, \\ 0, & inak. \end{cases}$$

a) Rozhodnite o úlohe symbolu a. Je to parameter, alebo konkrétna hodnota?
b) Určte $P(X < 0.5)$.
c) Určte c tak, aby $P(X < c) = 0.8$.

2.2.6 Nech funkcia hustoty f je daná vzťahom

$$f(x) = \begin{cases} ax(1-x), & 0 < x < 1, \\ 0, & inak. \end{cases}$$

a) Rozhodnite o úlohe symbolu a. Je to parameter, alebo konkrétna hodnota?
b) Určte $P(0.25 < X < 0.75)$.
c) Určte m tak, aby $P(X < m) = 0.5$.

2.2.7 Nech funkcia hustoty f je daná vzťahom

$$f(x) = \begin{cases} a(4-3x^2), & -1 < x < 1, \\ 0, & inak. \end{cases}$$

a) Rozhodnite o úlohe symbolu a. Je to parameter, alebo konkrétna hodnota?
b) Určte $P(|X| < 0.5)$.

2.2.8 Nech funkcia hustoty f je pre všetky $x \in R$ daná vzťahom

$$f(x) = \frac{a}{1+x^2}$$

a) Rozhodnite o úlohe symbolu a. Je to parameter, alebo konkrétna hodnota?
b) Určte $P(|X| < 1)$.
c) Určte m tak, aby $P(X < m) = 0.5$.

2.2.9 Nech funkcia hustoty f je daná vzťahom

$$f(x) = \begin{cases} axe^{-x^2/2}, & x > 0, \\ 0, & inak. \end{cases}$$

a) Rozhodnite o úlohe symbolu a. Je to parameter, alebo konkrétna hodnota?
b) Určite $P(X < 1)$.

2.2.10 Nech funkcia f je daná vzťahom

$$f(x) = \begin{cases} \frac{2}{a}\left(1-\frac{x}{a}\right), & 0 < x < a, \\ 0, & inak. \end{cases}$$

a) Ukážte, že pre každé kladné a je funkcia f hustota (a je teda parameter).
b) Nech X je náhodná veličina, ktorej rozdelenie určuje f. Nájdite $P(X < a/2)$.
c) Určite m tak, aby $P(X < m) = 0.5$.

2.2.11 Nech funkcia f je daná vzťahom

$$f(x) = \begin{cases} \frac{1}{2}(1+ax), & -1 < x < 1, \\ 0, & inak. \end{cases}$$

a) Ukážte, že pre každé $a \in \langle -1, 1\rangle$ je funkcia f hustota. Znázornite grafy f postupne pre hodnoty $a = -1, -0.5, 0, 0.5, 1$.
b) Nech X je náhodná veličina, ktorej rozdelenie určuje f. Nájdite $P(X < 0)$.

2.2.12 Nech funkcia f je daná vzťahom

$$f(x) = \begin{cases} \frac{1}{a}\left(1-\frac{|x|}{a}\right), & -a < x < a, \\ 0, & inak. \end{cases}$$

a) Ukážte, že pre ľubovoľné kladné a je funkcia f hustota (znázornite graf).
b) Určite $P(|X| < \frac{a}{2})$.
c) Určite c tak, aby platilo $P(X < c) = 0.9$.

2.2.13 Nech funkcia f je daná vzťahom

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & inak. \end{cases}$$

a) Ukážte, že pre ľubovoľné kladné λ je funkcia f hustota (λ je teda parametrom rozdelenia). Znázornite grafy f napr. pre $\lambda = 0.5$ a pre $\lambda = 5$. Všimnite si, že hustota f v prípade $\lambda = 5$, nadobúda v okolí 0 (napravo od nuly) hodnoty blízke hodnote 5.
b) Určite m tak, aby $P(X < m) = 0.5$.

2.3 Distribučná funkcia

Zopakujme, že náhodná veličina X je po pravdepodobnostnej stránke známa, keď poznáme jej *rozdelenie*. To znamená, že dokážeme stanoviť hodnoty $\mathrm{P}(a \le X < b)$, pre akékoľvek $a, b \in \mathrm{R}$. Inými slovami, vieme určiť pravdepodobnosť toho, že v pokuse sa X realizuje hodnotou, ktorá leží v intervale $\langle a, b)$.

Ak X je diskrétna veličina a $H(X) = \{x_1, x_2, \dots, x_n, \dots\}$, tak rozdelenie X určuje pravdepodobnostná funkcia f vzťahom

$$\mathrm{P}(a \le X < b) = \textstyle\sum_{i:\, a \le x_i < b} f(x_i)$$

Ak X je spojitá s hustotou, tak rozdelenie X určuje práve hustota f, pretože platí

$$\mathrm{P}(a \le X < b) = \textstyle\int_a^b f(x)\,dx.$$

V tomto článku sa oboznámime s distribučnou funkciou, ktorá jednoduchým a univerzálnym spôsobom určuje rozdelenie náhodnej veličiny tak pre diskrétne, ako aj pre spojité veličiny s hustotou (Veta 2.3.2). Vezmime na vedomie, že svet náhodných veličín je naozaj veľmi pestrý a tie veličiny, s ktorými pracujeme, totiž, diskrétne (na jednej strane) a spojité s hustotou (na strane druhej), sú len dva krajné prípady, ktoré ho zďaleka nevyčerpávajú. Distribučná funkcia je univerzálny pojem v tom zmysle, že popisuje rozdelenie akejkoľvek náhodnej veličiny. Jej dôležitosť vynikne v článku 2.5.

2.3.1 Definícia. Nech (Ω, S, P) je pravdepodobnostný priestor a X je náhodná veličina. Distribučná funkcia F je reálna funkcia reálnej premennej definovaná vzťahom

$$F\colon \mathrm{R} \to \langle 0, 1\rangle$$

$$F(x) = \mathrm{P}(\{\omega\colon X(\omega) < x\}) = \mathrm{P}(X < x).$$

2.3.2 Veta. Nech X je náhodná veličina a F je jej distribučná funkcia. Distribučná funkcia určuje rozdelenie veličiny X, pretože pre všetky $a, b \in \mathrm{R}$, $a < b$, platí

$$\mathrm{P}(a \le X < b) = F(b) - F(a).$$

Dôkaz. Ak $a, b \in \mathrm{R}$, $a < b$, tak pre udalosti zrejme platí: $\{X < a\} \cup \{a \le X < b\} = \{X < b\}$, pričom udalosti na ľavej strane rovnosti sú disjunktné. Z aditivity P máme

$$\mathrm{P}(X < a) + \mathrm{P}(a \le X < b) = \mathrm{P}(X < b),$$

čo znamená, že $F(a) + \mathrm{P}(a \le X < b) = F(b)$, a to je ekvivalentné dokazovanej rovnosti.

2.3.3 Veta. Nech X je náhodná veličina a F je jej distribučná funkcia. Potom platí:

1 F je neklesajúca na R, t. j. ak $a, b \in \mathrm{R}$, $a < b$, potom $F(a) \le F(b)$.

2 F je spojitá zľava v každom svojom argumente $x \in \mathrm{R}$.

3 $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$.

Dôkaz. Prvý záver plynie z vety 2.3.2, pretože $0 \le \mathrm{P}(a \le X < b) = F(b) - F(a)$, čo znamená, že $0 \le F(b) - F(a)$, teda $F(a) \le F(b)$. Druhý a tretí záver sú dôsledkom spojitosti P.

2.3.4 Poznámka. Dá sa dokázať, že každú funkciu $G\colon \mathrm{R} \to \langle 0, 1\rangle$, ktorá má vlastnosti 1, 2, 3 z vety 2.3.3, môžeme považovať za distribučnú funkciu nejakej náhodnej veličiny. Možno totiž zostrojiť (Ω, S, P) a definovať na ňom X tak, aby funkcia G bola distribučnou funkciou veličiny X. Ak teda zadanie úlohy začína slovami „nech F je distribučnou funkciou" (veličiny X), stačí overiť, že F má vlastnosti 1, 2, 3 a akceptujeme, že taká náhodná veličina X (ktorej F je distribučná funkcia) naozaj existuje (aj keď ju nemáme zadanú).

2.3.5 Veta. Nech X je diskrétna veličina s pravdepodobnostnou funkciou f a hodnotovou množinou $H(X) = \{x_1, x_2, x_3, \dots\}$. Potom pre jej distribučnú funkciu F platí

$$F(x) = \sum_{i:\, x_i < x} f(x_i),$$

pre všetky $x \in \mathrm{R}$. Ak X je spojitá s hustotou f, potom pre jej distribučnú funkciu F platí

$$F(x) = \int_{-\infty}^{x} f(t)dt,$$

pre všetky $x \in \mathrm{R}$, a preto F je spojitá funkcia (vo všetkých $x \in \mathrm{R}$). V tých bodoch x, v ktorých je f spojitá, má F deriváciu a platí:

$$F'(x) = f(x).$$

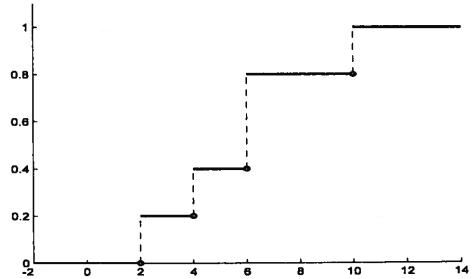
2.3.6 Príklad. Rozdelenie diskrétnej veličiny X je dané tabuľkou jej pravdepodobnostnej funkcie

x_i	2	4	6	10
$f(x_i)$	0.2	0.2	0.4	0.2

Nájdime predpis pre distribučnú funkciu F veličiny X a znázornime jej graf.

Riešenie. Využijeme vzťah z vety 2.3.5, $F(x) = \sum_{x_i < x} f(x_i)$.

Zrejme $F(x) = 0$, pre $x \le 2$, pretože niet takých hodnôt x_i, pre ktoré platí $x_i < x$ (ak $x \le 2$). Ak pre x platí $2 < x \le 4$, $F(x) = 0.2$, pretože iba hodnota $x_1 = 2$ leží naľavo od x. Ak však napríklad $x = 4.1$, tak sčítavame hodnoty f v argumentoch 2 a 4, pretože to sú práve tie, ktoré ležia naľavo od $x = 4.1$. Preto $F(4.1) = 0.2 + 0.2 = 0.4$. Hodnoty F v ostatných argumentoch sa dajú ľahko vyčítať z grafu na obr. 2-3.



Obr. 2-3. Graf distribučnej funkcie F z príkladu 2.3.6

Ako vidíme, pre body $x \in H(X) = \{2, 4, 6, 10\}$ platí: $\mathrm{P}(X = x) = F(x+0) - F(x)$, kde pod symbolom $F(x+0)$ rozumieme limitu sprava v bode x, t. j. $F(x+0) = \lim_{u \to x+} F(u)$.

Distribučná funkcia F diskrétnej náhodnej veličiny X je vždy po častiach konštantná, skoková funkcia. Výšky skokov sú hodnoty $\mathrm{P}(X = x_i)$, teda hodnoty $f(x_i)$.

2.3.7 Príklad. Rozdelenie spojitej veličiny X je dané hustotou f vzťahom

$$f(x) = \begin{cases} e^{-x}, & x > 0, \\ 0, & inak. \end{cases}$$

Graf funkcie f je na obr. 2-4a. Nájdime predpis pre distribučnú funkciu F.

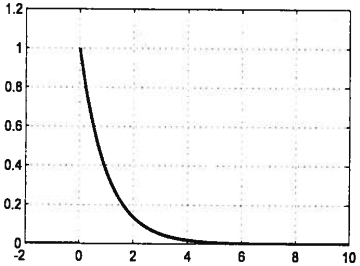
Riešenie. Postupujeme podľa vzťahu $F(x) = \int_{-\infty}^{x} f(t)\, dt$ (veta 2.3.5). Vzhľadom na predpis hustoty, výpočet realizujeme najprv pre záporné a potom pre kladné argumenty.

Pre $x \le 0$ zrejme $F(x) = 0$, pretože $F(x) = \int_{-\infty}^{x} f(x)\, dx = \int_{-\infty}^{x} 0\, dx = 0$.

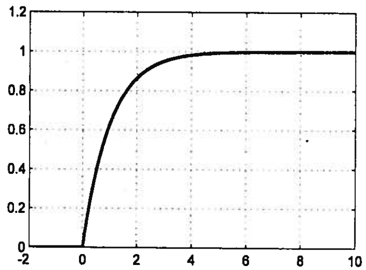
Pre kladné x máme $F(x) = \int_{-\infty}^{x} f(t)\, dt = \int_{-\infty}^{0} 0\, dt + \int_{0}^{x} e^{-t}\, dt = 0 + [-e^{-t}]_0^x = 1 - e^{-x}$.

Pre F teda platí

$$F(x) = \begin{cases} 1 - e^{-x}, & x > 0, \\ 0, & inak. \end{cases};$$



Obr. 2-4a. Graf funkcie hustoty



Obr. 2-4b. Graf distribučnej funkcie

Zopakujme, že distribučná funkcia veličiny s hustotou je vždy spojitá a má deriváciu v tých bodoch, v ktorých je hustota spojitá. V našom prípade F nemá deriváciu v bode $x = 0$.

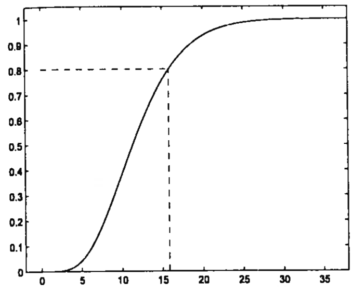
Veta 2.3.2 hovorí, že distribučná funkcia určuje rozdelenie náhodnej veličiny, a preto je prirodzené, že pomocou distribučnej funkcie je možné definovať dôležité charakteristiky rozdelenia. Medzi ne patrí pojem kvantilu, resp. p-kvantilu, ktorý budeme využívať v štatistike. Symbol p predstavuje parameter a je ním ľubovoľné číslo z intervalu (0, 1). V nasledujúcej definícii sa obmedzíme len na niektoré spojité rozdelenia s hustotou, keď definovanie pojmu kvantil je jednoduché (v štatistike s takými prípadmi vystačíme).

2.3.8 Definícia. Nech X má rozdelenie s hustotou f a pritom také, že distribučná funkcia F je na množine $H(X) = \{x \in \mathrm{R} : f(x) > 0\}$ rastúca. Nech $p \in (0, 1)$. Číslo x_p je p-kvantil rozdelenia veličiny X, teda rozdelenia, ktoré určuje F, ak platí

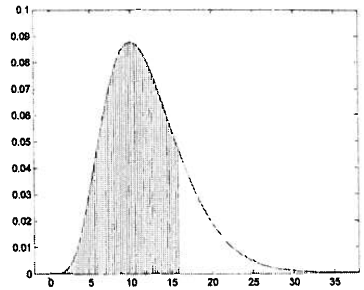
$$F(x_p) = p,$$

čo znamená, že $\mathrm{P}(X < x_p) = p$, resp. vyjadrené pomocou hustoty: $\int_{-\infty}^{x_p} f(x)dx = p$.

Na obrázku 2-5a je znázornený p-kvantil, pre $p = 0.8$, prostredníctvom distribučnej funkcie. Na obrázku 2-5b je rovnaký kvantil znázornený prostredníctvom funkcie hustoty.



Obr. 2-5a. Číslo 15.8 je 0.8-kvantil



Obr. 2-5b. $x_{0.8} = 15.8$, lebo $\int_{-\infty}^{15.8} f(x)dx = 0.8$

2.3.9 Príklad. Určime 0.9-kvantil rozdelenia z príkladu 2.3.7.

Riešenie. Pre kladné argumenty distribučná funkcia rozdelenia má predpis $F(x) = 1 - e^{-x}$. Podľa definície 0.9-kvantilu, pre $x_{0.9}$ má platiť

$$1 - e^{-x_{0.9}} = 0.9,$$

teda $e^{-x_{0.9}} = 0.1$, odkiaľ $-x_{0.9} = \ln(0.1)$, a teda $x_{0.9} = \ln(10)$. Overte, že $F(\ln 10) = 0.9$.

Úlohy.

2.3.1 Nech X je diskrétna veličina, ktorej pravdepodobnostná funkcia je daná tabuľkou

x_i	0	1	2	4	8
$f(x_i)$	0.1	0.2	0.4	0.2	0.1

Nájdite predpis pre distribučnú funkciu F tohoto rozdelenia, načrtnite jej graf a určte hodnoty $F(2)$, $F(3)$ a $F(8)$.

2.3.2 Uistite sa, že ku všetkým rozdeleniam z článku 2.2 viete nájsť príslušné distribučné funkcie. V každej situácii overte, že získaná funkcia F má vlastnosti uvedené vo vete 2.3.3.

2.3.3 Uistite sa, že ku všetkým rozdeleniam z článku 2.2 viete nájsť kvantily, t. j. viete určiť x_p tak, aby platilo: $F(x_p) = p$. V ktorých úlohách, teda pre ktoré rozdelenia, je problém riešiť rovnicu $F(x_p) = p$? V úlohách 2.2.6 a 2.2.7 ide o kubické rovnice (algebraické rovnice tretieho stupňa), ktoré vieme riešiť len numerickými metódami.

2.3.4 Funkcia F je pre všetky $x \in \mathrm{R}$ definovaná vzťahom

$$F(x) = \frac{1}{a + e^{-x}}$$

a) Určite hodnotu konštanty a tak, aby F bola distribučná funkcia, to znamená tak, aby mala všetky tri vlastnosti z vety 2.3.3.

b) Nájdite hustotu rozdelenia, ktoré určuje funkcia F.

c) Určte 0.5-kvantil tohoto rozdelenia.

2.4 Niektoré dôležité rozdelenia

V tomto článku sa oboznámime s rozdeleniami, ktoré sa v reálnych situáciách objavujú veľmi často. Vieme, že rozdelenie je určené buď distribučnou funkciou F, alebo hustotou f, resp. pravdepodobnostnou funkciou f. Zvolíme definovanie rozdelení prostredníctvom ich pravdepodobnostnej funkcie f, resp. hustoty f, pretože sa s nimi pracuje pohodlnejšie. Okrem toho vplyv hodnoty parametrov sa na grafe funkcie f prejavuje zreteľnejšie ako na grafe distribučnej funkcie. Začneme s diskrétnymi rozdeleniami.

2.4.1 Hovoríme, že X má *alternatívne rozdelenie* s parametrom p, symbolicky: $X \sim \mathrm{A}(p)$, ak pravdepodobnostná funkcia f je daná tabuľkou

x	0	1
$f(x)$	$1-p$,	p

Vidíme, že množina možných hodnôt má iba dva body, $H(X) = \{0, 1\}$. Ide teda o najjednoduchšiu náhodnú veličinu (jednoduchšou je iba taká, ktorá má jedinú hodnotu, avšak taká už fakticky nie je náhodná). Význam parametra p je zrejmý, $p = \mathrm{P}(X=1)$, teda $p \in (0, 1)$. Situácií, v ktorých stretávame toto rozdelenie, je bezpočet, napr.

- V pokuse, ktorý spočíva v hode mincou, keď jednej strane mince prisúdime hodnotu 1 a druhej strane hodnotu 0. Ak je minca symetrická a vyvážená, $p = 0.5$, v opačnom prípade mincu považujeme za falošnú a p je nejaký iný bod intervalu (0, 1).
- V pokuse s hracou kockou, v ktorom X uvažujeme ako *indikátor* padnutia čísla 6. Ak je kocka symetrická a homogénna, $p = 1/6$, inak ju považujeme za falošnú a hodnotu parametra p považujeme buď za známy, alebo za neznámy bod intervalu (0, 1). Samozrejme, X môže byť indikátor aj iného čísla, dokonca, X môže byť indikátor nejakej udalosti, napr. padnutia čísla menšieho ako 5. V takom prípade – ak je kocka normálna – zrejme $X \sim \mathrm{A}(p)$, kde $p = 2/3$.
- Všimnime si, že v akomkoľvek náhodnom pokuse môžeme zvoliť akúkoľvek náhodnú udalosť A a mať na mysli veličinu X, ako indikátor udalosti A. Teda X nadobúda v pokuse hodnotu 1 práve vtedy, ak A nastane (a zrejme X sa rovná nule práve vtedy, ak A nenastane). Platí: $X \sim \mathrm{A}(p)$, kde $p = \mathrm{P}(\mathrm{A})$.

2.4.2 Hovoríme, že X má *rovnomerné rozdelenie* na množine $\{x_1, x_2, \ldots, x_n\}$, symbolickým zápisom: $X \sim \mathrm{R}\{x_1, x_2, \ldots, x_n\}$, ak pravdepodobnostná funkcia f je daná tabuľkou

x_i	x_1	x_2	...	x_n
$f(x_i)$	$\frac{1}{n}$	$\frac{1}{n}$	...	$\frac{1}{n}$

Zrejme v pokuse s bežnou hracou kockou máme do činenia s $X \sim \mathrm{R}\{1, 2, \ldots, 6\}$, alebo pri hre s ruletou ide o rozdelenie $\mathrm{R}\{0, 1, 2, \ldots, 36\}$. Aj keď v týchto prípadoch sú hodnoty x_i celé čísla, v zásade $x_1, x_2, \ldots, x_n$ môžu byť akékoľvek reálne čísla a chápeme ich ako parametre tohoto rozdelenia. Tak napr. do osudia môžeme dať 10 lístkov a na každý z nich napísať nejaké iné reálne číslo. Ak náhodne ťaháme lístok z osudia a za výsledok pokusu považujeme číslo na vytiahnutom lístku, tak ide o realizovanie náhodnej veličiny X s rovnomerným rozdelením (na tých desiatich číslach x_i, ktoré sme na lístky napísali).

Sofistikovanejšie realizovanie pokusu s týmto rozdelením umožňujú generátory náhodných čísel, či už na počítači (v nejakom výpočtovom prostredí), alebo hoci len na vedeckej kalkulačke. Funkcie RND, *random*, či *rand* poskytujú náhodné čísla z intervalu (0, 1) a ak z nich použijeme, napríklad, len prvú cifru zľava za desatinnou bodkou, tak získame generátor rovnomerného rozdelenia na množine $\{0, 1, \ldots, 9\}$.

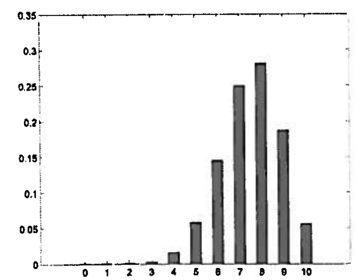
2.4.3 Hovoríme, že X má *binomické rozdelenie* s parametrami n, p (n je prirodzené číslo, p je reálne číslo z intervalu (0, 1)) a píšeme $X \sim \mathrm{Bi}(n, p)$, ak pravdepodobnostná funkcia f je daná predpisom

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x \in \{0, 1, 2, \ldots, n\}, \\ 0, & inde. \end{cases}$$

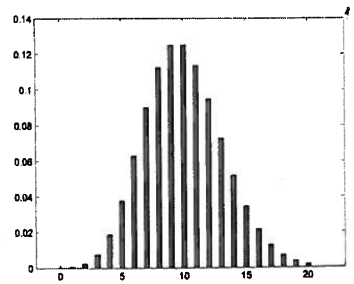
To znamená, že $H(X) = \{0, 1, 2, \ldots, n\}$ a pre $k \in \{0, 1, 2, \ldots, n\}$, $f(k) = \binom{n}{k} p^k (1-p)^{n-k}$.

Prvý raz sme sa s takýmto rozdelením stretli v článku 1.7, v odseku 1.7.6. Išlo o zložený pokus, ktorý spočíval v n-násobnom opakovaní jednoduchého pokusu, v ktorom sme sledovali nastatie akejsi udalosti A. V i-tom opakovaní pokusu indikátorom udalosti A je veličina X_i, a tak súčet $X_1 + X_2 + \ldots + X_n$ predstavuje počet prípadov, keď v sérii n-opakovaní nastane udalosť A. Podľa vzťahu (2) odseku 1.7.6 súčet $X_1 + X_2 + \ldots + X_n$ má binomické rozdelenie s parametrami n, p. Spomeňme konkrétne situácie, v ktorých išlo o binomické rozdelenie:

- Príklad 1.7.7, desaťnásobný hod mincou, veličina X predstavovala počet znakov v sérii 10 hodov. Zrejme $X \sim \mathrm{Bi}(n, p)$, kde $n = 10$, $p = 0.5$.
- Úloha 1.7.3, desaťnásobný hod mincou, teda opäť $n = 10$, ale teraz falošnou mincou, keď pravdepodobnosť padnutia znaku $p = 0.55$. Ak X znamená počet padnutých znakov, tak X má rozdelenie Bi(10, 0.55).
- Príklad 1.7.8, kde veličina X predstavuje počet nepodarkov spomedzi 100 vytiahnutých výrobkov, náhodným výberom s vrátením. Zrejme $X \sim \mathrm{Bi}(n, p)$, kde $n = 100$ a p je pravdepodobnosť vyrobenia nepodarku.
- Úloha 1.7.7, v ktorej X znamená počet hrozienok v zakúpenom koláčiku. Riešenie úlohy spočívalo v prijatí hypotézy: $X \sim \mathrm{Bi}(n, p)$, pričom $n = 10\,000$, $p = 0.001$.
- Binomické rozdelenie sme použili aj v úlohách 1.7.8 až 1.7.12.



Obr. 2-6a. Hodnoty f pre Bi(10, 0.75)



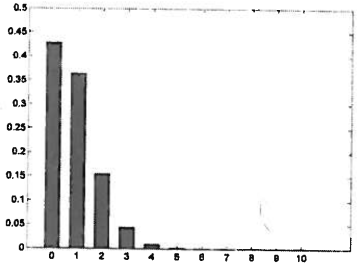
Obr. 2-6b. Hodnoty f pre Bi(10 000, 0.001)

Diagramy obr. 2-6 vizualizujú tabuľky pravdepodobnostných funkcií binomického rozdelenia. Na obr. 2-6a, kde ide o Bi(10, 0.75), sú hodnoty $f(0)$ a $f(1)$ také malé, že sa povedľa ostatných hodnôt nedajú zobraziť ($f(0) = 9.5\,10^{-7}$ a $f(1) = 2.9\,10^{-5}$). Obr. 2-6b vizualizuje rozdelenie Bi(10 000, 0.001). Od hodnoty 21 sa hodnoty f nezobrazujú, napr. $f(21) \approx 9.10^{-4}$.

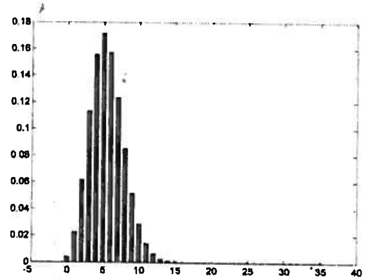
2.4.4 Hovoríme, že X má *Poissonovo rozdelenie* s parametrom λ ($\lambda > 0$), píšeme $X \sim \mathrm{Po}(\lambda)$, ak pravdepodobnostná funkcia je daná vzťahom

$$f(x) = \begin{cases} e^{-\lambda}\dfrac{\lambda^x}{x!}, & x \in \{0, 1, 2, \dots, k, \dots\}, \\ 0, & \textit{inak.} \end{cases}$$

To znamená, že $H(X) = \{0, 1, 2, \dots\}$ a pre $k \in \{0, 1, 2, \dots\}$, $f(k) = e^{-\lambda}\dfrac{\lambda^k}{k!}$. Predstavu o tvare rozdelenia dávajú obrázky 2-7, ktoré zobrazujú hodnoty pravdepodobnostnej funkcie rozdelenia pre $\lambda = 0.85$, resp. pre $\lambda = 5.5$.



Obr. 2-7a. $X \sim \mathrm{Po}(\lambda)$, $\lambda = 0.85$



Obr. 2-7b. $X \sim \mathrm{Po}(\lambda)$, $\lambda = 5.5$

Kde máme do činenia s týmto rozdelením? Napríklad tam, kde náhodnú veličinu je možné považovať za veličinu s binomickým rozdelením, pričom n je „veľké" a p „malé". Čo je veľké a čo je malé? Samozrejme, to nemožno jednoznačne povedať, ale ak n je rádu 10^2, resp. 10^3 a súčin np sú rádovo jednotky, resp. desiatky, tak je možné overiť, že platí:

$$\text{Ak } X \sim \mathrm{Bi}(n, p), \text{ tak } \mathrm{P}(X = k) \approx \mathrm{P}(Y = k), \text{ pre } Y \sim \mathrm{Po}(\lambda), \text{ kde } \lambda = np.$$

Napr. v príklade 1.7.8, nech pravdepodobnosť vyrobenia nepodarku, $p = 0.02$. Keď X znamená počet nepodarkov spomedzi 100 vytiahnutých výrobkov (výberom s vrátením), tak podľa odseku 2.4.2 vieme, že $X \sim \mathrm{Bi}(n, p)$, kde $n = 100$ a $p = 0.02$. Tabuľka pravdepodobnostnej funkcie f veličiny X je takáto:

x_i	0	1	2	3	4	5	6	...
$f(x_i)$	0.1326	0.2707	0.2734	0.1823	0.0902	0.0353	0.0114	...

Teraz tabelujme pravdepodobnostnú funkciu g veličiny Y, $Y \sim \mathrm{Po}(\lambda)$, pričom $\lambda = np = 2$.

y_i	0	1	2	3	4	5	6	...
$g(y_i)$	0.1353	0.2707	0.2707	0.1804	0.0902	0.0361	0.0120	...

Ako vidíme, zhoda nie je úplná, ale je veľmi dobrá. Je to dôsledok tvrdenia, ktorý je známy ako Poissonova veta.

2.4.5 Veta. (Poissonova veta) Nech $X_n \sim \mathrm{Bi}(n, p_n)$, $p_n \in (0, 1)$, tak, že $\lim_{n\to\infty} n p_n = \lambda$ a nech $Y \sim \mathrm{Po}(\lambda)$. Potom

$$\lim_{n\to\infty} \mathrm{P}(X_n = k) = \mathrm{P}(Y = k),$$

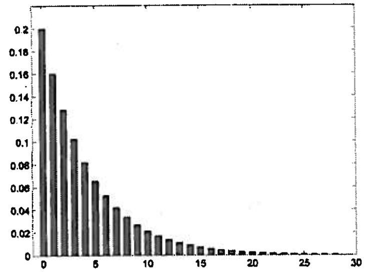
t. j.

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1 - p_n)^{n-k} = e^{-\lambda}\frac{\lambda^k}{k!}.$$

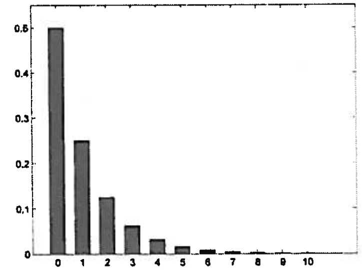
2.4.6 Hovoríme, že X má *geometrické rozdelenie* s parametrom p, $p \in (0, 1)$, $X \sim \mathrm{G}(p)$, ak pravdepodobnostná funkcia je daná vzťahom

$$f(x) = \begin{cases} (1-p)^x p, & \textit{ak } x \in \{0, 1, 2, \dots, k, \dots\}, \\ 0, & \textit{inak.} \end{cases}$$

čo znamená, že $H(X) = \{0, 1, 2, \dots\}$ a pre $k \in \{0, 1, 2, \dots\}$, $f(k) = (1-p)^k p$. Pre predstavu, obrázky 2-8 zobrazujú hodnoty pravdepodobnostnej funkcie pre $p = 0.2$, resp. $p = 0.5$.



Obr. 2-8a. $X \sim \mathrm{G}(p)$, $p = 0.2$



Obr. 2-8b. $X \sim \mathrm{G}(p)$, $p = 0.5$

Jednoduchú situáciu, v ktorej vystupuje rozdelenie $\mathrm{G}(p)$, $p = 0.5$, predstavuje pokus, v ktorom hádžeme normálnou mincou dovtedy, kým sa nedočkáme padnutia znaku. Nech X znamená počet „neúspešných" hodov, teda počet hodov, v ktorých znak nepadol (padnutím znaku pokus končí). Keď hádžeme kockou, kým nepadne číslo 6, tak počet neúspešných hodov je veličina s rozdelením $\mathrm{G}(p)$, $p = 1/6$. Voľne môžeme povedať, že geometrické rozdelenie modeluje v zloženom pokuse čas čakania na nejakú udalosť (čas meriame diskrétne), keď pravdepodobnosť tej udalosti (v jednom opakovaní) sa rovná p.

Pomenovanie *geometrické* rozdelenie je prirodzené, pretože overovanie toho, že súčet

$$f(0) + f(1) + f(2) + \dots = p + (1-p)p + (1-p)^2 p + \dots = p(1 + (1-p) + (1-p)^2 + \cdots)$$

sa rovná 1, vedie – ako vidíme – na súčet geometrického radu s kvocientom $1 - p$. Všimnime si, že hodnotu distribučnej funkcie $F(n)$, pre $n > 0$, môžeme vyjadriť takto:

$$F(n) = \mathrm{P}(X < n) = \mathrm{P}(X = 0) + \mathrm{P}(X = 1) + \dots + \mathrm{P}(X = n-1) = p + (1-p)p + \dots + (1-p)^{n-1}p =$$

$$= p(1 + (1-p) + (1-p)^2 + \dots + (1-p)^{n-1}) = p\,\frac{1-(1-p)^n}{1-(1-p)} = 1 - (1-p)^n.$$

Preto pre $\mathrm{P}(X \ge n)$ máme

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

Teraz už ľahko ukážeme, že platí

$$P(X \geq m + k \mid X \geq m) = P(X \geq k),$$

pretože

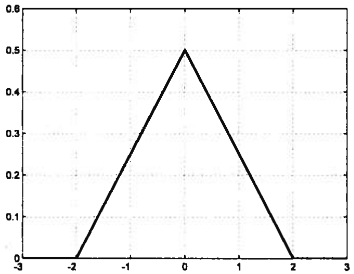
$$P(X \geq m + k \mid X \geq m) = \frac{P(X \geq m + k)}{P(X \geq m)} = \frac{(1-p)^{m+k}}{(1-p)^m} = (1-p)^k = P(X \geq k).$$

Hovoríme, že geometrické rozdelenie je „bez pamäte". Čo znamená táto vlastnosť v pokuse s mincou, ktorú hádžeme dovtedy, kým nepadne znak? Ak v troch hodoch znak nepadol, tak pravdepodobnosť toho, že ďalšie dva hody neskončia znakom je práve taká veľká, ako pravdepodobnosť toho, že keď s hádzaním začíname, tak prvé dva hody budú neúspešné (vzali sme $m = 3$, $k = 2$). Skrátka, minca si, samozrejme, minulosť nepamätá.

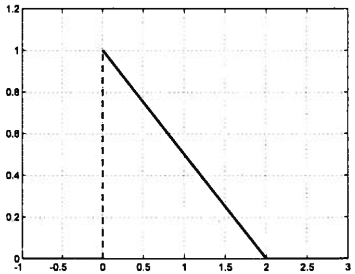
2.4.7 Teraz predstavíme niekoľko spojitých rozdelení. Prvým bude *trojuholníkové* rozdelenie s parametrami a, b, c ($a \leq b \leq c$, $a \neq c$). Píšeme $X \sim \Delta(a, b, c)$, ak hustota X je daná vzťahom

$$f(x) = \begin{cases} \dfrac{2(x-a)}{(b-a)(c-a)}, & a \leq x \leq b, \\ \dfrac{2(c-x)}{(c-a)(c-b)}, & b \leq x \leq c, \\ 0, & \textit{inde}. \end{cases}$$

Graf hustoty vytvára trojuholník (obrázky 2-9).



Obr. 2-9a. $X \sim \Delta(-2, 0, 2)$



Obr. 2-9b. $X \sim \Delta(0, 0, 2)$

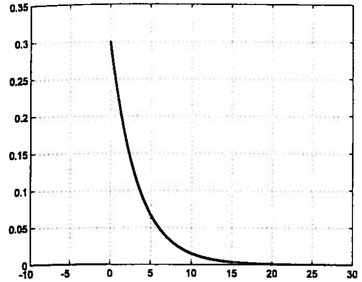
2.4.8 Hovoríme, že X má *rovnomerné* rozdelenie na intervale (a, b), píšeme $X \sim R(a, b)$, ak hustotou X je funkcia

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & ak\ x \in (a, b), \\ 0, & \textit{inak}. \end{cases}$$

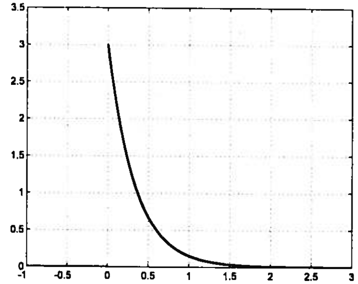
Hustota je teda na intervale (a, b) konštantná. Jej jednoduchý graf nezobrazujeme, zrejme vykresľuje obdĺžnik. Preto sa niekedy hovorí o rovnomernom rozdelení ako o *obdĺžnikovom* rozdelení.

2.4.9 Hovoríme, že X má *exponenciálne* rozdelenie s parametrom λ ($\lambda > 0$), $X \sim \text{Exp}(\lambda)$, ak hustota je daná vzťahom

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$



Obr. 2-10a. Graf hustoty Exp(λ), $\lambda = 0.3$



Obr. 2-10b. Graf hustoty Exp(λ), $\lambda = 3$

Nech s, t sú kladné reálne čísla. Ukážme, že pre exponenciálne rozdelenie platí

$$P(X \geq s + t \mid X \geq s) = P(X \geq t),$$

čo znamená (pozri odsek 2.4.6), že exponenciálne rozdelenie je rozdelením „bez pamäte". Najprv uvážme, že pre distribučnú funkciu tohto rozdelenia, pre $x > 0$, platí:

$$F(x) = 1 - e^{-\lambda x},$$

(čo dostaneme analogicky, ako v 2.3.7), a preto $P(X \geq t) = 1 - P(X < t) = e^{-\lambda t}$. Potom však naozaj

$$P(X \geq s + t \mid X \geq s) = \frac{P(X \geq s + t)}{P(X \geq s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X \geq t).$$

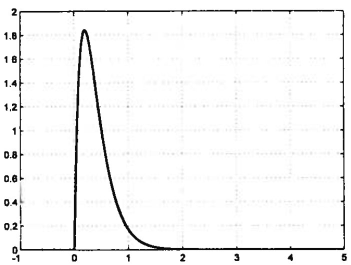
Pripomeňme, že v odseku 2.4.6 sme ukázali, že medzi diskrétnymi rozdeleniami takúto vlastnosť má geometrické rozdelenie.

2.4.10 Hovoríme, že X má *Erlangovo* rozdelenie s parametrami k, λ ($k \in \mathbb{N}$, $\lambda > 0$), čo budeme zapisovať $X \sim \text{Erl}(k, \lambda)$, ak hustota rozdelenia je daná vzťahom

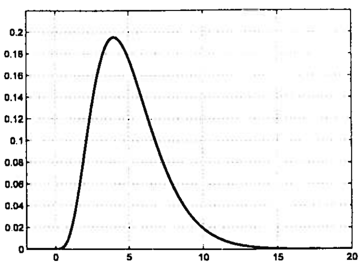
$$f(x) = \begin{cases} \dfrac{\lambda^k}{(k-1)!} x^{k-1} e^{-\lambda x}, & ak\ x > 0, \\ 0, & \textit{inak}. \end{cases}$$

Všimnime si, že Erlangovo rozdelenie s parametrom $k = 1$, je exponenciálne rozdelenie. V kapitole 3 sa dozvieme, že súčet nezávislých náhodných veličín s exponenciálnym rozdelením je veličinou s Erlangovým rozdelením. Obe tieto rozdelenia sa veľmi často objavujú v úlohách a problémoch teórie hromadnej obsluhy.

Predstavu o tvare funkcií hustôt dávajú obr. 2-11a, 2-11b, kde sú znázornené grafy hustôt Erlangovho rozdelenia pre dve dvojice parametrov: $(k, \lambda) = (2, 5)$, resp. $(k, \lambda) = (5, 1)$.



Obr. 2-11a. Graf hustoty Erl(2, 5)

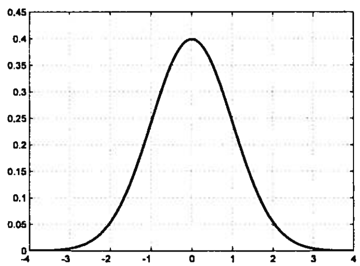


Obr. 2-11b. Graf hustoty Erl(5, 1)

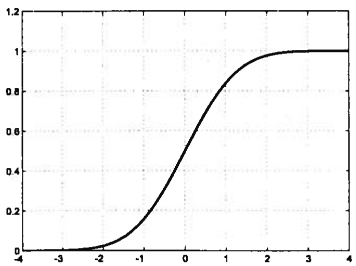
2.4.11 Najdôležitejším rozdelením je nepochybne *normálne* rozdelenie N(μ, σ^2), nazývané tiež *Gaussovo* rozdelenie. Parameter μ je ľubovoľné reálne číslo, parameter σ je kladné reálne číslo (neprehliadnime, že v zápise N(μ, σ^2), – v zhode s konvenciou – uvádzame druhú mocninu σ^2). Budeme písať $X \sim$ N(μ, σ^2), ak hustota rozdelenia je daná vzťahom

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x\in(-\infty,\infty).$$

Parameter μ predstavuje bod (na osi x), v ktorom hustota dosahuje globálne maximum. Graf hustoty je symetrický podľa vertikály $x = \mu$. Preto je parameter μ označovaný ako parameter polohy. Parameter σ je parameter tvaru, lebo ovplyvňuje štíhlosť krivky hustoty. Čím je σ menšie, tým je Gaussova krivka štíhlejšia. Body $\mu - \sigma$ a $\mu + \sigma$, sú inflexné body. Medzi nimi je hustota konkávna, inde je konvexná. Neskôr sa dozvieme, že parameter μ je strednou hodnotou rozdelenia a σ^2 je jeho varianciou (resp. σ je smerodajnou odchýlkou). Najprv sa venujme významnému prípadu, keď $\mu = 0$ a $\sigma = 1$. Rozdelenie N(0, 1) nazývame *štandardné* (resp. *normované*) normálne rozdelenie. Vľavo je graf jeho hustoty f_N a vpravo graf jeho distribučnej funkcie F_N (označenia f_N a F_N budeme v ďalšom dôsledne používať).



Obr. 2-12a. Graf hustoty N(0, 1)



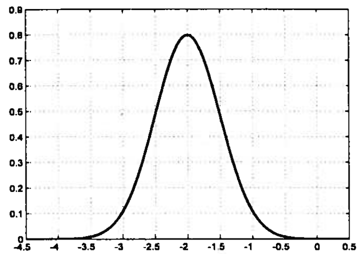
Obr. 2-12b. Distribučná funkcia N(0, 1)

Dôležitým faktom je, že distribučnú funkciu normálneho rozdelenia nie je možné napísať explicitne tak, ako to bolo možné urobiť napr. v prípade exponenciálneho rozdelenia. Dôvodom je, že primitívna funkcia k funkcii $f(x) = exp(-x^2)$ sa nedá vyjadriť ako kombinácia elementárnych funkcií, a preto na integrovanie hustoty normálneho rozdelenia

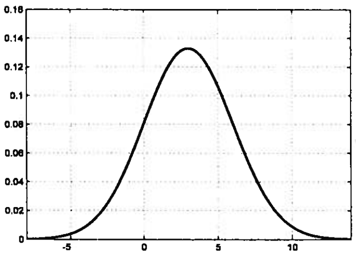
nemožno aplikovať Leibnitzov-Newtonov vzorec. Musíme sa zmieriť s tým, že distribučnú funkciu vyjadríme ako funkciu, ktorej argument je horná hranica integrálu, t. j. v normovanom prípade v tvare $F_N(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{\frac{-t^2}{2}}\, dt$, resp. vo všeobecnom prípade v tvare

$$F(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\, dt.$$

V článku 2.5 ukážeme, ako sa k hodnotám $F(x;\mu,\sigma^2)$ dostaneme bez integrovania. Využijeme to, že hodnoty distribučnej funkcie F_N rozdelenia N(0, 1) sú numericky spočítané a uvedené v tabuľkách. Aj keď dnes (vďaka počítačom a výpočtovým prostrediam) už význam štatistických tabuliek nie je taký veľký ako v minulom storočí, stále štandardné postupy s tabuľkami patria do základnej výbavy každého, kto chce rozumieť základom pravdepodobnosti a štatistiky. Obr. 2-13 ilustrujú vplyv parametrov na tvar grafu hustoty.



Obr. 2-13a. Graf hustoty N(-2, 0.25)



Obr. 2-13b. Graf hustoty N(3, 9)

Úlohy

2.4.1 Nech $X \sim \Delta(-2, 0, 2)$, $Y \sim \Delta(0, 0, 2)$. Nájdite predpis pre ich distribučné funkcie a znázornite ich grafy.

2.4.2 Nech $X \sim$ R(a, b). Určte predpis pre distribučnú funkciu veličiny X a načrtnite jej graf.

2.4.3 Nech $X \sim$ Exp(λ), $\lambda = 2$. Znázornite graf distribučnej funkcie F veličiny X. Určte $x_{0,9}$, t. j. nájdite hodnotu 0.9-kvantilu rozdelenia Exp(2).

2.4.4 Nech $X \sim$ N(0, 1). Ukážte, že z párnosti hustoty f_N pre distribučnú funkciu F_N platí:

$$\text{Ak } x < 0, \text{ tak } F_N(x) = 1 - F_N(-x).$$

Vďaka tejto skutočnosti vystačíme s tabelovaním $F_N(x)$ pre kladné argumenty x.

2.4.5 Nech $X \sim$ N(0, 1). Aplikujte výsledok z predchádzajúcej úlohy a určte

a) $P(-1 < X < 2)$, b) $P(-2 < X < 0)$,
c) $P(-0.5 < X)$, d) $P(-2 < X < 1)$.

2.4.6 Pomocou tabuliek overte, že pre $X \sim$ N(0, 1) platí:

$$P(-1 < X < 1) = 0.68268, \text{ t. j. } P(-1 < X < 1) \approx 0.683,$$
$$P(-2 < X < 2) = 0.95450, \text{ t. j. } P(-2 < X < 2) \approx 0.955,$$
$$P(-3 < X < 3) = 0.99730, \text{ t. j. } P(-3 < X < 3) \approx 0.997.$$

2.5 Rozdelenie transformovanej náhodnej veličiny

Aj keď náhodná veličina je pre nás zobrazenie, pracujeme s ňou trochu špecifickým spôsobom. To, čo nás eminentne zaujíma, je jej rozdelenie. Z článku 2.3 vieme, že rozdelenie náhodnej veličiny určuje distribučná funkcia F vzťahom

$$P(a \le X < b) = F(b) - F(a), \text{ pre všetky } a, b \in R,$$

ktorý má univerzálnu platnosť. Platí tak pre diskrétne veličiny, ako aj pre spojité s hustotou a aj pre všetky ostatné (v úlohe 2.5.20 jednu takú veličinu predstavíme). Dôležitosť distribučnej funkcie vynikne práve v tomto článku, v ktorom pôjde o určenie rozdelenia novej náhodnej veličiny Y, ktorá je transformáciou pôvodnej veličiny X, t. j. $Y = h(X)$.

Vstupom úlohy je rozdelenie veličiny X a transformácia, ktorú predstavuje daná funkcia h. Cieľom snaženia je určiť rozdelenie veličiny $Y = h(X)$. Rozdelenie veličiny X je dané buď funkciou f, alebo F. Keďže pracujeme s dvoma náhodnými veličinami X, Y, označme ich distribučné funkcie F_X a F_Y. Dané sú teda funkcia F_X a funkcia h. Úlohou je určiť F_Y.

Najčastejšie transformácie sú lineárna funkcia, $h(x) = a + bx$ a jednoduchá kvadratická, $h(x) = x^2$. Principiálne však funkciou h môže byť akákoľvek reálna funkcia reálnej premennej. Samozrejme, definičný obor funkcie h musí obsahovať obor hodnôt veličiny X (táto podmienka zaručuje, že veličina $Y = h(X)$ je dobre definovaná).

V prvých troch situáciách ide o prípady, keď veličina X je diskrétna, a preto aj veličina Y, $Y = h(X)$, je diskrétna. Pri hľadaní pravdepodobnostnej funkcie f_Y veličiny Y v odsekoch 2.5.1 a 2.5.2 vystačíme s aditivitou pravdepodobnosti, pretože $H(X)$ je konečná.

2.5.1 Príklad. Nech rozdelenie veličiny X je dané pravdepodobnostnou funkciou vo forme tabuľky

x_i	-2	-1	0	1	2
$f_X(x_i)$	0.1	0.2	0.3	0.2	0.2

Nájdime rozdelenie náhodných veličín $Y = 1 + 2X$, $Z = X^2$.

Riešenie. Ak f_Y je pravdepodobnostná funkcia veličiny Y, tak máme

y_i	-3	-1	1	3	5
$f_Y(y_i)$	0.1	0.2	0.3	0.2	0.2

Pretože napr. $f_Y(-3) = P(Y = -3) = P(1 + 2X = -3) = P(X = -2) = f_X(-2) = 0.1$. Analogicky, pre f_Z

z_i	0	1	4
$f_Z(z_i)$	0.3	0.4	0.3

Napr. $f_Z(4) = P(Z = 4) = P(X^2 = 4) = P(\{X = -2\} \cup \{X = 2\}) = f_X(-2) + f_X(2) = 0.1 + 0.2 = 0.3$.

2.5.2 Príklad. Nech $X \sim Bi(6, 0.4)$. Určime rozdelenie veličín $Y = (X-3)^2$, $Z = \sin\left(\frac{\pi}{2}X\right)$.

Riešenie. Najprv uveďme tabuľku pravdepodobnostnej funkcie veličiny X.

x_i	0	1	2	3	4	5	6
$f_X(x_i)$	0.0467	0.1866	0.3110	0.2765	0.1382	0.0369	0.0041

Zrejme hodnotami pre $Y = (X-3)^2$ sú čísla 9, 4, 1 a 0. Napr. pre $P(Y = 9)$ máme

$$P(Y = 9) = P((X-3)^2 = 9) = P(\{X - 3 = -3\} \cup \{X - 3 = 3\}) = P(X = 0) + P(X = 6) = 0.0508.$$

$P(Y = 4)$ stanovme menej formálne; udalosť $\{Y = 4\}$ nastane práve vtedy, keď nastane udalosť $\{X = 1\}$, alebo udalosť $\{X = 5\}$. Preto

$$P(Y = 4) = P(X = 1) + P(X = 5) = 0.1866 + 0.0369 = 0.2235.$$

Analogicky,

$$P(Y = 1) = P(X = 2) + P(X = 4) = 0.3110 + 0.1382 = 0.4492.$$

Nakoniec, $P(Y = 0) = P(X = 3) = 0.2765$, pretože udalosť $\{Y = 0\}$ nastane práve vtedy, keď nastane udalosť $\{X = 3\}$. Takto pre f_Y máme

y_i	0	1	4	9
$f_Y(y_i)$	0.2765	0.4492	0.2235	0.0508

Rovnako postupujeme v prípade $Z = \sin\left(\frac{\pi}{2}X\right)$. Pravdepodobnostná funkcia veličiny Z je daná tabuľkou

z_i	-1	0	1
$f_Z(z_i)$	0.2765	0.50	0.2235

Napr. $f_Z(0) = P(Z = 0) = P(X = 0) + P(X = 2) + P(X = 4) + P(X = 6) = 0.50$. Doteraz sme využívali len aditivitu pravdepodobnosti. V nasledujúcom príklade musíme uplatniť silnejšiu vlastnosť – sigma aditivitu.

2.5.3 Príklad. Nech $X \sim G(p)$ a $Y = min(X, 4)$. Nájdime rozdelenie náhodnej veličiny Y.

Riešenie. Pripomeňme, že ak $X \sim G(p)$, tak $H(X) = \{0, 1, 2, \dots, n, \dots\}$ a $P(X = k) = (1-p)^k p$. $Y = min(X, 4)$ sa rovná nule práve vtedy, keď $X = 0$. Formálne, $\{Y = 0\} = \{X = 0\}$, a preto

$$P(Y = 0) = P(X = 0) = (1-p)^0 p = p.$$

Rovnako pre $k = 1, 2, 3$ máme $\{Y = k\} = \{X = k\}$, a tak $P(Y = k) = (1-p)^k p$. Avšak pre $k = 4$ máme

$$P(Y = 4) = P(X = 4) + P(X = 5) + P(X = 6) + \dots = (1-p)^4 p + (1-p)^5 p + (1-p)^6 p + \dots =$$

$$= (1-p)^4 p\,[\,1 + (1-p) + (1-p)^2 + \dots\,] = (1-p)^4 p\,[\,\frac{1}{1-(1-p)}\,] = (1-p)^4.$$

Pre pravdepodobnostnú funkciu veličiny Y dostávame

$$f_Y(k) = \begin{cases} (1-p)^k p, & k = 0, 1, 2, 3, \\ (1-p)^4, & k = 4. \end{cases}$$

Prehľadnejšie formou tabuľky

k	0	1	2	3	4
$f_Y(k)$	p	$(1-p)p$	$(1-p)^2 p$	$(1-p)^3 p$	$(1-p)^4$

Nasledujúca veta formuluje vzťahy medzi rozdeleniami X a Y vo všeobecnej rovine.

2.5.4 Veta. Nech rozdelenie diskrétnej veličiny X je dané pravdepodobnostnou funkciou f. Nech h je reálna funkcia reálnej premennej a taká, že veličina $Y = h(X)$ je dobre definovaná,

t. j. obor hodnôt veličiny X je podmnožinou definičného oboru funkcie h. Potom pre pravdepodobnostnú funkciu g veličiny Y platí

$$g(y) = \sum_{x:\, h(x)=y} f(x).$$

Dôkaz.

$$g(y) = \mathrm{P}(Y = y) = \mathrm{P}\left(\bigcup_{x:\, h(x)=y} \{X = x\}\right) = \sum_{x:\, h(x)=y} \mathrm{P}(X = x) = \sum_{x:\, h(x)=y} f(x).$$

Teraz sa venujme transformáciám veličín, ktorých rozdelenie určuje hustota.

2.5.5 Príklad. Nech $X \sim \mathrm{R}(0, 2)$ a $Y = 1 + 2X$. Určime rozdelenie veličiny Y.

Riešenie. Intuitívne očakávame, že $2X$ má rozdelenie $\mathrm{R}(0, 4)$, pretože rovnomerné rozdelenie $\mathrm{R}(0, 2)$ sa len „roztiahne" na rovnomerné rozdelenie na $(0, 4)$. Veličina $1 + 2X$ je len „posunutie" veličiny $2X$, takže $1 + 2X$ má rozdelenie $\mathrm{R}(1, 5)$. Toľko intuícia a teraz presne. Začneme odvodením vzťahu medzi distribučnými funkciami veličín Y a X. Neznámu F_Y chceme vyjadriť pomocou známej F_X

$$F_Y(y) = \mathrm{P}(Y < y) = \mathrm{P}(1 + 2X < y) = \mathrm{P}(2X < y - 1) = \mathrm{P}(X < \tfrac{y-1}{2}) = F_X\left(\tfrac{y-1}{2}\right).$$

Funkcia F_X je daná vzťahom

$$F_X(x) = \begin{cases} 0, & x \le 0, \\ x/2, & 0 < x \le 2, \\ 1, & x > 2. \end{cases}$$

To znamená, že

$$F_Y(y) = F_X\left(\frac{y-1}{2}\right) = \begin{cases} 0, & \dfrac{y-1}{2} \le 0, \\ \dfrac{y-1}{4}, & 0 < \dfrac{y-1}{2} \le 2, \\ 1, & \dfrac{y-1}{2} > 2, \end{cases}$$

resp.

$$F_Y(y) = \begin{cases} 0, & y \le 1, \\ \dfrac{y-1}{4}, & 1 < y \le 5, \\ 1, & y > 5. \end{cases}$$

Ako vidíme, F_Y je lineárna na intervale $(1, 5)$, $F_Y(1) = 0$, $F_Y(5) = 1$, teda veličina Y má rovnomerné rozdelenie na intervale $(1, 5)$, čo je v zhode s tým, čo sme intuitívne očakávali.

2.5.6 Príklad. Nech $X \sim \Delta(0, 2, 2)$, $Y = 3 + \frac{X}{2}$. Nájdime rozdelenie veličiny Y.

Riešenie. Najprv argumentujme intuitívne. Ak $X \sim \Delta(0, 2, 2)$, tak veličina $\frac{X}{2}$ má rozdelenie $\Delta(0, 1, 1)$ a keďže veličina $Y = 3 + \frac{X}{2}$ je len posunutím veličiny $\frac{X}{2}$, náš tip je: $Y \sim \Delta(3, 4, 4)$.

Teraz odvoďme rozdelenie veličiny Y presným postupom. Opäť začneme odvodením vzťahu medzi distribučnými funkciami veličín X a Y. Funkcia F_X je daná vzťahom

$$F_X(x) = \begin{cases} 0, & x \le 0, \\ x^2/4, & 0 < x \le 2, \\ 1, & x > 2 \end{cases}$$

a pre distribučnú funkciu veličiny Y platí

$$F_Y(y) = \mathrm{P}(Y < y) = \mathrm{P}(3 + \tfrac{X}{2} < y) = \mathrm{P}(\tfrac{X}{2} < y - 3) = \mathrm{P}(X < 2(y - 3)) = F_X(2y - 6).$$

Preto

$$F_Y(y) = F_X(2y - 6) = \begin{cases} 0, & 2y - 6 \le 0, \\ (2y - 6)^2/4, & 0 < 2y - 6 \le 2, \\ 1, & 2y - 6 > 2. \end{cases}$$

Po úprave, $F_Y(y) = 0$, pre $y \le 3$, resp. $F_Y(y) = (y - 3)^2$, pre $y \in (3, 4)$, resp. $F_Y(y) = 1$, pre $y > 4$. Derivovaním funkcie F_Y získame hustotu g. Zisťujeme, že pre $y \in (3, 4)$ dostávame $g(y) = 2y - 6$, a to znamená, že $Y = 3 + \frac{X}{2} \sim \Delta(3, 4, 4)$, čo je v zhode s našou intuíciou.

2.5.7 Veta. (O lineárnej transformácii normálneho rozdelenia)

1. Ak $X \sim \mathrm{N}(0, 1)$ a $b \ne 0$, tak $bX \sim \mathrm{N}(0, b^2)$ a $Y = a + bX \sim \mathrm{N}(a, b^2)$.
2. Ak $X \sim \mathrm{N}(\mu, \sigma^2)$ a $Y = \frac{X - \mu}{\sigma}$, tak $Y \sim \mathrm{N}(0, 1)$.
3. Ak $X \sim \mathrm{N}(\mu, \sigma^2)$ a $Y = a + bX$, $(b \ne 0)$, tak $Y \sim \mathrm{N}(a + b\mu, b^2\sigma^2)$.

Dôkaz. Nech $X \sim \mathrm{N}(0, 1)$. Ukážeme, že $Y = a + bX \sim \mathrm{N}(a, b^2)$. Začneme odvodením distribučnej funkcie veličiny Y. Ak b je kladné, tak

$$F_Y(y) = \mathrm{P}(Y < y) = \mathrm{P}(a + bX < y) = \mathrm{P}(bX < y - a) = \mathrm{P}(X < \tfrac{y-a}{b}) = F_X(\tfrac{y-a}{b}) = F_N(\tfrac{y-a}{b}),$$

pretože $F_X = F_N$. Keďže distribučnú funkciu normálneho rozdelenia nemožno písať celkom explicitne (len v tvare funkcie hornej hranice integrálu z hustoty), hľadajme hneď hustotu g veličiny Y, ktorú získame derivovaním F_Y

$$g(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_N(\tfrac{y-a}{b}) = f_N(\tfrac{y-a}{b})\, \frac{d}{dy}(\tfrac{y-a}{b}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\frac{y-a}{b})^2}{2}} \frac{1}{b} = \frac{1}{\sqrt{2\pi}\, b} e^{-\frac{(y-a)^2}{2b^2}}.$$

Funkcia g je zrejme hustota rozdelenia $\mathrm{N}(a, b^2)$, preto $Y \sim \mathrm{N}(a, b^2)$. Ak b je záporné, tak

$$F_Y(y) = \mathrm{P}(a + bX < y) = \mathrm{P}(bX < y - a) = \mathrm{P}(X > \tfrac{y-a}{b}) = 1 - \mathrm{P}(X < \tfrac{y-a}{b}) = 1 - F_N(\tfrac{y-a}{b}).$$

Derivovaním F_Y získame hustotu rovnako, ako v prípade $b > 0$. Ostatné časti vety sa dokážu analogicky.

Veta 2.5.7 má zásadný význam a často sa budeme na ňu odvolávať. Umožňuje vyjadriť hodnoty $F(x; \mu, \sigma^2)$ prostredníctvom funkcie F_N, teda pomocou distribučnej funkcie normovaného normálneho rozdelenia. To ilustruje nasledujúci príklad.

2.5.8 Príklad. Nech $X \sim \mathrm{N}(3, 4)$. Určime $\mathrm{P}(1 \le X < 6)$ využitím tabuliek funkcie F_N.

Riešenie. Zrejme $\mathrm{P}(1 \le X < 6) = F_X(6) - F_X(1)$, ale tabelovanú máme iba F_N. Využijeme však bod 2 vety 2.5.7

$$\mathrm{P}(1 \le X < 6) = \mathrm{P}(1 - 3 \le X - 3 < 6 - 3) = \mathrm{P}\left(\tfrac{1-3}{2} \le \tfrac{X-3}{2} < \tfrac{6-3}{2}\right) = F_N(1.5) - F_N(-1) =$$
$$= F_N(1.5) - [\,1 - F_N(1)\,] = 0.93319 + 0.84134 - 1 = 0.77453.$$

2.5.9 Príklad. Nech $X \sim R(-2, 2)$ a $Y = X^2$. Určime rozdelenie veličiny Y.

Riešenie. Ak $X \sim R(-2, 2)$, tak $H(X) = (-2, 2)$ a pre $Y = X^2$ máme $H(Y) = \langle 0, 4)$. Preto pre F_Y zrejme platí: $F_Y(y) = 0$, pre $y \leq 0$ a $F_Y(y) = 1$, pre $y \geq 4$. Vzťah medzi F_Y a F_X stačí hľadať pre argumenty $y \in (0, 4)$. Ak $0 < y < 4$, tak

$$F_Y(y) = P(Y < y) = P(X^2 < y) = P(-\sqrt{y} < X < \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}).$$

Pripomeňme, že

$$F_X(x) = \begin{cases} 0, & x \leq -2, \\ \dfrac{x+2}{4}, & -2 < x \leq 2, \\ 1, & x > 2. \end{cases}$$

Keďže pre $y \in (0, 4)$ argumenty $\sqrt{y}$, resp. $-\sqrt{y}$ ležia v intervale $(-2, 2)$, pre $F_Y(y)$ máme

$$F_Y(y) = F_X(\sqrt{y}) - F_X(-\sqrt{y}) = \frac{\sqrt{y}+2}{4} - \frac{-\sqrt{y}+2}{4} = \frac{\sqrt{y}+\sqrt{y}}{4} = \frac{\sqrt{y}}{2}$$

Predpis pre hustotu g veličiny Y na $(0, 4)$ získame derivovaním odvodenej funkcie F_Y. Zrejme mimo intervalu $(0, 4)$ sa hustota g rovná nule. Takto dostávame

$$g(y) = \begin{cases} \dfrac{1}{4\sqrt{y}}, & 0 < y < 4, \\ 0, & inak. \end{cases}$$

Načrtnite grafy odvodených funkcií F_Y, resp. g. Všimnite si, že hustota g je v pravom okolí bodu 0 neohraničená. Overte, že integrál funkcie g na intervale $(0, 4)$ sa rovná 1.

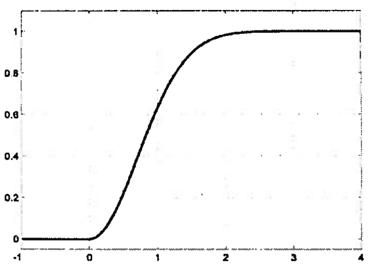
2.5.10 Príklad. Nech $X \sim Exp(\lambda)$ a $Y = \sqrt{\lambda X}$. Určime rozdelenie veličiny Y.

Riešenie. Pretože $H(X) = (0, \infty)$ a λ je kladné reálne číslo, je zrejme $H(Y) = (0, \infty)$. Preto stačí hľadať vzťah medzi F_Y a F_X pre kladné argumenty. Ak $y > 0$, potom

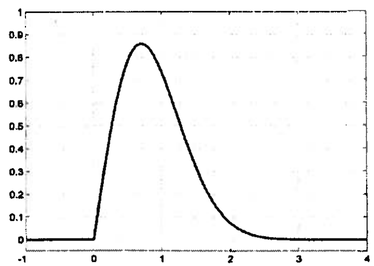
$$F_Y(y) = P(Y < y) = P(\sqrt{\lambda X} < y) = P(\lambda X < y^2) = P(X < \frac{y^2}{\lambda}) = F_X(\frac{y^2}{\lambda}).$$

Vieme, že pre $x > 0$ $F_X(x) = 1 - e^{-\lambda x}$, a preto pre $y > 0$ máme $F_Y(y) = 1 - e^{-\lambda \frac{y^2}{\lambda}} = 1 - e^{-y^2}$. Hustotu g veličiny Y dostaneme derivovaním F_Y,

$$F_Y(y) = \begin{cases} 0, & y \leq 0, \\ 1 - e^{-y^2}, & y > 0, \end{cases} \qquad g(y) = \begin{cases} 0, & y \leq 0, \\ 2y\, e^{-y^2}, & y > 0. \end{cases}$$



Obr. 2-14 a. Distribučná funkcia z príkladu 2.5.10



Obr. 2-14 b. Hustota g z príkladu 2.5.10

2.5.11 Príklad. Nech $X \sim N(0, 1)$ a $Y = X^2$. Nájdime rozdelenie veličiny Y.

Riešenie. Pretože veličina $X \sim N(0, 1)$, namiesto F_X píšeme F_N. Ako sme to už viackrát urobili, začneme odvádzaním distribučnej funkcie F_Y veličiny Y. Je zrejmé, že $H(X) = (-\infty, \infty)$, ale $H(Y) = \langle 0, \infty)$. Preto nech $y > 0$. Pre $F_Y(y)$ máme

$$F_Y(y) = P(Y < y) = P(X^2 < y) = P(-\sqrt{y} < X < \sqrt{y}) = F_N(\sqrt{y}) - F_N(-\sqrt{y}).$$

Pretože distribučnú funkciu F_N rozdelenia $N(0, 1)$ vieme písať len v tvare funkcie hornej medze integrálu z hustoty, postupujeme ďalej hľadaním hustoty g veličiny Y derivovaním.

$$g(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy}\left(F_N(\sqrt{y}) - F_N(-\sqrt{y})\right) = f_N(\sqrt{y})\frac{1}{2\sqrt{y}} + f_N(-\sqrt{y})\frac{1}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{y}}\, e^{-\frac{y}{2}}$$

Hustota g je takto daná vzťahom

$$g(y) = \begin{cases} 0, & y \leq 0, \\ \dfrac{1}{\sqrt{2\pi}}\dfrac{1}{\sqrt{y}}\, e^{-\frac{y}{2}} & y > 0. \end{cases}$$

Všimnime si, že funkcia g je v pravom okolí bodu 0 neohraničená.

Úlohy

2.5.1 Nech rozdelenie diskrétnej veličiny X je dané tabuľkou. Určte rozdelenie veličín
a) $Y = 7 - 2X$, b) $Z = (X - 2)^2$,
c) $U = \sin(\pi X)$, d) $V = \cos(\pi X)$.

x_i	0	1	2	4	6
$f(x_i)$	0.1	0.2	0.4	0.2	0.1

2.5.2 Nech $X \sim R\{-2, -1, 0, 1, 2\}$. Nájdite rozdelenie veličiny
a) $Y = X^2$, b) $Z = X^3$.

2.5.3 Nech $X \sim R\{1, 2, \ldots, 2n - 1\}$. Nájdite rozdelenie veličiny
a) $Y = (X - n)^2$, b) $Z = (X - n)^3$.

2.5.4 Nech $X \sim G(p)$. Nájdite rozdelenie veličiny
a) $Y = max(X, 5)$, b) $Z = \frac{1}{X+1}$.

2.5.5 Nech $X \sim Po(\lambda)$. Nájdite rozdelenie veličiny
a) $Y = X^2$, b) $Z = \sin(\frac{\pi X}{2})$.

2.5.6 Nech $X \sim N(0, 1)$. Určte rozdelenie veličiny
a) $Y = 3 + 2X$, b) $Z = 2 - 3X$.

2.5.7 Nech $X \sim N(2, 9)$. Určte $P(-1 \leq X \leq 8)$. Aké rozdelenie má veličina $Y = \frac{X-2}{3}$?

2.5.8 Nech $X \sim N(3, 16)$. Určte a, b tak, aby veličina $Y = a + bX$ mala rozdelenie $N(0, 1)$.

2.5.9 Nech $X \sim N(4, 4)$ a $Y = 3 - \frac{X}{2}$. Určte $P(-0.5 < Y < 2)$.

2.5.10 Nech $X \sim N(-3, 9)$. Určte rozdelenie veličiny
a) $Y = 7 - 2X$, b) $Z = 4 + \frac{X}{3}$.

2.5.11 Nech $X \sim R(0, 4)$. Nájdite rozdelenie veličiny
a) $Y = X^2$, b) $Z = \sqrt{X}$.

2.5.12 Nech $X \sim R(0, c)$, $c > 0$. Nájdite rozdelenie veličiny
a) $Y = X + a$, b) $Z = X^2$.

2.5.13 Nech $X \sim \Delta(-1, 0, 1)$. Nájdite rozdelenie veličiny
a) $Y = 2X$, b) $Z = X + 1$, c) $U = X^2$.

2.5.14 Nech $X \sim \Delta(0, 1, 1)$. Nájdite rozdelenie veličiny $Y = X^2$.

2.5.15 Nech $X \sim \Delta(0, a, 2a)$. Nájdite rozdelenie veličiny
a) $Y = X - a$, b) $Z = (X - a)^2$.

2.5.16 Ukážte, že platí: Ak $X \sim R(0, 1)$, tak $\sqrt{X} \sim \Delta(0, 1, 1)$.

2.5.17 Nech $X \sim \text{Exp}(\lambda)$. Nájdite rozdelenie veličiny
a) $Y = cX$, $c > 0$, b) $Z = \sqrt{X}$, c) $U = X + 1$, d) $V = X^2$.

2.5.18 Nech X má hustotu danú vzťahom $f(x) = x + 0.5$, pre $x \in (0, 1)$, $f(x) = 0$, inde. Nájdite rozdelenie veličiny
a) $Y = X + 1$, b) $Z = 2X$, c) $U = X^2$.

2.5.19 Nech X má spojitú distribučnú funkciu F, ktorá je na intervale $\langle a, b)$ rastúca, pričom platí: $F(a) = 0$, $F(b) = 1$ (ak $b = \infty$, tak $F(b)$ znamená limitu $F(x)$, pre $x \to \infty$). Označme zúženie funkcie F na interval $\langle a, b)$ symbolom $\tilde{F}$. Nech $F^{-1} : \langle 0, 1) \to \langle a, b)$ je inverznou funkciou k zúženiu $\tilde{F}$. Platí:

$$\tilde{F} : \langle a, b) \to \langle 0, 1), \; F^{-1} : \langle 0, 1) \to \langle a, b) \text{ a zrejme } F(F^{-1}(u)) = u, \text{ pre } u \in \langle 0, 1).$$

Nech $U \sim R(0, 1)$. Ukážte, že veličina $Y = F^{-1}(U)$ má rovnaké rozdelenie ako X, t. j. $F_Y = F$.

Poznámka. Každé výpočtové prostredie má generátor náhodných čísel, t. j. dokáže generovať (simulovať) realizácie veličiny $U \sim R(0, 1)$. Vďaka uvedenému tvrdeniu ľahko generujeme realizácie náhodnej veličiny X (ktorej distribučná funkcia je F), pokiaľ dokážeme nájsť inverziu F^{-1}.

Za realizácie veličiny X považujeme hodnoty $Y = F^{-1}(U)$. Tak napríklad, keď chceme generovať hodnoty z rozdelenia $\text{Exp}(\lambda)$, položíme $Y = F^{-1}(U) = -\ln(1 - U)/\lambda$.

2.5.20 Nech $X \sim R(0, 3)$ a funkcia h je po častiach lineárna: $h(x) = 0$, pre $x < 1$, $h(x) = x - 1$, pre $x \in \langle 1, 2)$, $h(x) = 1$, pre $x > 2$. Nech $Y = h(X)$. Ukážte, že veličina Y nie je diskrétna. Avšak Y nie je ani spojitá s hustotou (prečo?). Y je však legitímnou náhodnou veličinou – nájdite jej distribučnú funkciu a načrtnite jej graf.

3 NÁHODNÉ VEKTORY

3.1 Náhodný vektor a jeho rozdelenie

Ak v rámci náhodného pokusu sledujeme súčasne viac ako jednu náhodnú veličinu, napr. dve, X a Y, tak výsledkom pokusu nie je jedno číslo $X(\omega)$, ale dvojica čísel $(X(\omega), Y(\omega))$. Hovoríme, že máme do činenia s dvojrozmerným náhodným vektorom (X, Y). Uvažujme o jednoduchej situácii, keď v hode dvoma hracími kockami X predstavuje minimum z bodov, ktoré padli na kockách a Y nech je súčet bodov. Samozrejme, obe veličiny sú nám dobre známe, avšak keď ich uvažujeme súčasne, v rámci toho istého pokusu (t. j. v rámci toho istého hodu), tak sa vynárajú zaujímavé otázky, súvisiace s väzbou (vzťahom) medzi X a Y. Napríklad, ak v pokuse sa X realizuje hodnotou 5, tak síce nevieme s istotou povedať, akou hodnotou sa realizovala Y, ale je celkom isté, že v tom pokuse nastala udalosť $\{Y \geq 10\}$. Cítime, že veličiny X, Y sú – voľne povedané – do istej miery závislé (presné definície sú predmetom tejto kapitoly).

Teraz vezmime celkom praktickú situáciu. Automat nituje výrobok na troch miestach a vykonáva zvary na štyroch miestach (toho istého výrobku). Nech X označuje počet nekvalitných nitov a Y počet nekvalitných zvarov na výrobku. Informáciu o kvalite výroby nesie dvojica veličín (X, Y). Je pochopiteľné, že nás zaujíma, či existuje nejaká (stochastická) väzba medzi X a Y. A ak tam väzba je, chceli by sme ju opísať, resp. kvantifikovať.

Budeme rozlišovať medzi vektorom (X, Y) ako zobrazením z Ω do R^2 a jeho realizáciou, ktorou je konkrétny bod (x, y) v rovine, $x = X(\omega)$, $y = Y(\omega)$. Aj keď náhodná veličina X je pre nás zobrazenie, t. j. $X: \Omega \to R$, v kapitole 2 sme si uvedomili, že viac ako zobrazeniu samotnému, venovali sme sa *rozdeleniu* náhodnej veličiny. To znamená, že sme hovorili o pravdepodobnostiach $P(a \leq X < b)$, pre $a, b \in R$. Analogická situácia bude aj v prípade náhodného vektora. Nakoniec vezmime na vedomie, že úvahy, ktoré urobíme pre prípad dvojrozmerného vektora, sa dajú prirodzene zovšeobecniť na prípad trojrozmerného, resp. viacrozmerného vektora.

3.1.1 Definícia. Nech X, Y sú náhodné veličiny na (Ω, S, P). Náhodný vektor $T = (X, Y)$ je usporiadaná dvojica veličín X, Y, ktorú chápeme ako zobrazenie T z Ω do $R \times R$

$$T: \Omega \to R \times R$$

$$T(\omega) = (X(\omega), Y(\omega)), \quad \text{pre všetky } \omega \in \Omega.$$

Pod *rozdelením* vektora (X, Y) rozumieme systém pravdepodobností $P(a \leq X < b, c \leq Y < d)$, pre všetky $a, b, c, d \in R$. Distribučná funkcia vektora (X, Y) je funkcia

$$F : R \times R \to \langle 0, 1\rangle$$

definovaná vzťahom

$$F(x, y) = P(X < x, Y < y), \text{ pre všetky } x, y \in R.$$

3.1.2 Poznámka. Zápis $P(X < x, Y < y)$ je zaužívaná skratka pre $P(\{X < x\} \cap \{Y < y\})$, teda pre pravdepodobnosť *súčasného* nastatia udalostí $\{X < x\}$, $\{Y < y\}$. Analogicky,

$$P(a \le X < b, c \le Y < d) \quad \text{znamená} \quad P(\{a \le X < b\} \cap \{c \le Y < d\}).$$

Rozdelenie vektora (X, Y) určuje distribučná funkcia, pretože pre všetky $a, b, c, d \in \mathrm{R}$ platí

$$P(a \le X < b, c \le Y < d) = F(b, d) - F(a, d) - F(b, c) + F(a, c).$$

Vzťah nie je ťažké dokázať, stačí opakovane využiť aditivitu pravdepodobnosti.

Distribučná funkcia vektora (X, Y) má síce zásadný význam pre teóriu, avšak ak sú obe zložky vektora diskrétne veličiny, ďaleko pohodlnejšie sa pracuje s pravdepodobnostnou funkciou vektora. Ukážeme, že aj pravdepodobnostná funkcia určuje rozdelenie náhodného vektora (bod c vety 3.1.4).

3.1.3 Definícia. Náhodný vektor (X, Y) je diskrétny, ak obe zložky X, Y sú diskrétne náhodné veličiny. Pravdepodobnostná funkcia $f(.,.)$ diskrétneho vektora (X, Y) je funkcia

$$f: \mathrm{R} \times \mathrm{R} \to \langle 0, 1 \rangle$$

definovaná vzťahom

$$f(x, y) = P(X = x, Y = y), \quad \text{pre všetky } x, y \in \mathrm{R}.$$

Pod množinou hodnôt vektora rozumieme tie body $(x, y) \in \mathrm{R} \times \mathrm{R}$, v ktorých $f(x, y) > 0$.

3.1.4 Veta. Nech $f(.,.)$ je pravdepodobnostná funkcia diskrétneho vektora (X, Y), ktorého zložky majú konečne veľa hodnôt, t. j. $H(X) = \{x_1, x_2, \ldots, x_m\}$ a $H(Y) = \{y_1, y_2, \ldots, y_n\}$. Ak $f_1(.)$, $f_2(.)$ sú pravdepodobnostné funkcie zložiek, tak platí

a) pre všetky $x, y \in \mathrm{R}$, $f(x, y) \ge 0$ a $\sum_{i=1}^{m} \sum_{j=1}^{n} f(x_i, y_j) = 1$,

b) pre všetky $x, y \in \mathrm{R}$, $f_1(x) = \sum_{j=1}^{n} f(x, y_j)$, $f_2(y) = \sum_{i=1}^{m} f(x_i, y)$,

c) pre všetky $a, b, c, d \in \mathrm{R}$, $P(a \le X < b, c \le Y < d) = \sum_{i: a \le x_i < b} \sum_{j: c \le y_j < d} f(x_i, y_j)$,
teda $f(.,.)$ určuje rozdelenie (X, Y).

Pri dokazovaní týchto tvrdení stačí využiť aditivitu pravdepodobnosti, lebo veta hovorí o vektoroch, ktorých zložky majú len konečne veľa hodnôt. V takom prípade budeme hodnoty funkcie $f(.,.)$ zapisovať v tvare dvojrozmernej tabuľky.

3.1.5 Príklad. Rub a líc mince označme nulou, resp. jednotkou. Hoďme mincou päťkrát a v pokuse definujme X, resp. Y takto: X je súčtom padnutých čísel z prvých troch hodov a Y nech predstavuje súčet padnutých čísel zo všetkých piatich hodov.

a) Nájdime pravdepodobnostnú funkciu $f(.,.)$ vektora (X, Y).
b) Určime pravdepodobnostné funkcie zložiek X, Y.

Riešenie. V tejto situácii je ľahké vidieť (X, Y) ako zobrazenie, keď Ω definujeme ako všetky usporiadané pätice núl a jednotiek. Ak napr. $\omega = (0, 1, 0, 1, 1)$, hodnotou vektora (X, Y) je dvojica $(1, 3)$, pretože $X(\omega) = 1$ a $Y(\omega) = 3$. Ak $\omega = (1, 0, 1, 1, 1)$, hodnotou vektora (X, Y) je dvojica $(2, 4)$, pretože $X(\omega) = 2$ a $Y(\omega) = 4$. Priestor Ω má zrejme 32 ($= 2^5$) bodov.

Pravdepodobnostnú funkciu je praktické zapisovať v tvare tabuľky:

$X \backslash Y$	0	1	2	3	4	5	f_1
0	$\frac{1}{32}$	$\frac{2}{32}$	$\frac{1}{32}$	0	0	0	$\frac{4}{32}$
1	0	$\frac{3}{32}$	$\frac{6}{32}$	$\frac{3}{32}$	0	0	$\frac{12}{32}$
2	0	0	$\frac{3}{32}$	$\frac{6}{32}$	$\frac{3}{32}$	0	$\frac{12}{32}$
3	0	0	0	$\frac{1}{32}$	$\frac{2}{32}$	$\frac{1}{32}$	$\frac{4}{32}$
f_2	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{10}{32}$	$\frac{10}{32}$	$\frac{5}{32}$	$\frac{1}{32}$	$\frac{32}{32}$

Veličina X má 4 možné hodnoty, Y ich má 6, ale vektor (X, Y) ich nemá $4 \cdot 6 = 24$, ale iba 12. Ako sme určili 12 kladných hodnôt $f(x_i, y_j)$? Napríklad $f(2, 3)$ takto:

$$P(X = 2, Y = 3) = P(\{(1,1,0,1,0), (1,1,0,0,1), (1,0,1,1,0), (1,0,1,0,1), (0,1,1,1,0), (0,1,1,0,1)\}) = \frac{6}{32}.$$

Pravdepodobnostné funkcie zložiek f_1, f_2 sme získali súčtami v riadkoch, resp. v stĺpcoch tabuľky, podľa bodu (b) vety 3.1.4. Ich hodnoty sa zvyknú uvádzať na okraji tabuľky. Aj preto sa funkciám f_1, f_2 hovorí *marginálne* pravdepodobnostné funkcie.

Analogicky ako v kapitole 2, vektor (X, Y) považujeme za definovaný, ak poznáme jeho rozdelenie (t. j. teraz, v diskrétnom prípade, poznáme tabuľku jeho pravdepodobnostnej funkcie f). Nasledujúci príklad ilustruje, že odpoveď na základné pravdepodobnostné otázky (napr. aká je $P(X < Y)$?) dostaneme z dvojrozmernej tabuľky funkcie f.

3.1.6 Príklad. Ľavá tabuľka predstavuje pravdepodobnostnú funkciu f vektora (X, Y).

$X \backslash Y$	1	2	3	4	
1	.1	.1	.05	0	
2	.05	.2	.1	.05	
3	0	.05	.2	.1	

$X \backslash Y$	1	2	3	4	f_1
1	.1	.1	.05	0	.25
2	.05	.2	.1	.05	.40
3	0	.05	.2	.1	.35
f_2	.15	.35	.35	.15	

a) Nájdime pravdepodobnostné funkcie zložiek.
b) Určime $P(1 \le X < 3, 2 \le Y < 4)$.
c) Určime $P(X < Y)$.

Riešenie. Podľa bodu (b) vety 3.1.4, sčítaním v riadkoch, resp. v stĺpcoch, tabuľky získame rozdelenia zložiek. Tak vznikli okraje v tabuľke vpravo. Ďalej

$$P(1 \le X < 3, 2 \le Y < 4) = P((X, Y) \in \{(1, 2), (1, 3), (2, 2), (2, 3)\}) = .1 + .05 + .2 + .1 = 0.45,$$

$$P(X < Y) = P((X, Y) \in \{(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4)\}) = .1 + \ldots + .05 + .1 = 0.40.$$

Nasledujúci príklad je venovaný reálnej situácii, v ktorej pravdepodobnostnú funkciu možno len aproximovať na základe dát, ktoré získame v sérii nezávislých opakovaní toho istého pokusu.

3.1.7 Príklad. Výrobný automat vykoná na blatníku auta dva zvary a nituje ho na troch miestach. Označme X počet nekvalitných zvarov a Y počet nekvalitných nitov na blatníku. Zrejme $H(X) = \{0, 1, 2\}$, $H(Y) = \{0, 1, 2, 3\}$. Sledovali sme výrobu 1000 blatníkov a na každom z nich sme zistili hodnoty X a Y. Výsledky uvádza tabuľka početností, ktorej sa hovorí *kontingenčná tabuľka*. Na každom objekte, v našom prípade na každom blatníku, zisťujeme súčasne hodnotu dvoch veličín X a Y.

$X\backslash Y$	0	1	2	3	
0	842	28	21	9	900
1	58	12	7	3	80
2	10	5	4	1	20
	910	45	32	13	1000

$X\backslash Y$	0	1	2	3	
0	.842	.028	.021	.009	.900
1	.058	.012	.007	.003	.080
2	.010	.005	.004	.001	.020
	.910	.045	.032	.013	1.0

Napríklad, v bunke odpovedajúcej prípadu $(X, Y) = (0, 0)$, číslo 842 predstavuje počet blatníkov (spomedzi 1000 preverovaných), v ktorých sme nezistili nekvalitný zvar, ani nekvalitný nit. Teda v 842 prípadoch (spomedzi 1000) sme pozorovali hodnotu $(0, 0)$. Bunky kontingenčnej tabuľky predstavujú výsledky pozorovania, teda empirické údaje.

Pravú tabuľku relatívnych početností považujeme za aproximáciu pravdepodobnostnej funkcie f vektora (X, Y). Tak napríklad, pravdepodobnosť $P(X > 1, Y > 1)$ aproximujeme hodnotou 0.005 $(= 0.004 + 0.001)$.

Doteraz sme pracovali s náhodnými vektormi, ktorých obe zložky boli diskrétne náhodné veličiny. Vieme si iste predstaviť náhodný vektor, ktorého obe zložky sú spojitej povahy. O relatívne jednoduchú situáciu ide vtedy, keď rozdelenie vektora určuje hustota. Na rozdiel od kapitoly 2, kde hustota bola funkcia jednej premennej, teraz hustota bude funkcia dvoch premenných.

3.1.8 Definícia. Náhodný vektor (X, Y) je spojitý vektor s hustotou $f(.,.)$, ak existuje integrovateľná funkcia $f: R \times R \to \langle 0, \infty)$ taká, že pre všetky $a, b, c, d \in R$, $a < b$, $c < d$ platí

$$P(a \le X < b, c \le Y < d) = \int_a^b \int_c^d f(x, y)\, dy\, dx.$$

Nasledujúca veta formuluje dôsledky uvedenej definície. Na rozdiel od diskrétneho prípadu, kde vystačíme s aditivitou, teraz v dôkaze je nutné využiť aj spojitosť pravdepodobnosti.

3.1.9 Veta. Nech (X, Y) je spojitý vektor s hustotou f. Potom platí

a) pre všetky $x, y \in R$ $f(x, y) \ge 0$ a $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$,

b) pre všetky $x, y \in R$ $F(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u, v)\, du\, dv$,

c) v bodoch $(x, y) \in R$, v ktorých je $f(.,.)$ spojitá, platí $f(x, y) = \frac{\partial^2 F}{\partial x \partial y}(x, y)$,

d) funkcie $f_1(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$, $f_2(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$ sú hustotami zložiek X, resp. Y,

e) ak A je merateľná množina v $R \times R$, tak $P((X, Y) \in A) = \iint_A f(x, y)\, dxdy$.

3.1.10 Príklad. Nech náhodný vektor (X, Y) má hustotu f danú vzťahom

$$f(x, y) = \begin{cases} ke^{-2x-3y}, & x > 0,\ y > 0 \\ 0, & inak. \end{cases}$$

Nájdime

a) hodnotu konštanty k,

b) hodnoty $F(3, 0)$ a $F(2, 1)$,

c) hustoty zložiek X, Y,

d) $P((X, Y) \in A)$, kde A je polrovina $A = \{(x, y): x + y < 1\}$.

Riešenie.

a) Pretože $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\, dx\, dy = k\int_0^{\infty}\int_0^{\infty} e^{-2x-3y} dy\, dx = k\frac{1}{6}$, podľa 3.1.9(a), zrejme $k = 6$.

b) Pre $F(3,0)$ máme $F(3, 0) = \int_{-\infty}^{3}\int_{-\infty}^{0} f(x, y)\, dy\, dx = \int_0^3\int_0^0 0\, dy\, dx = 0$. Analogicky platí

$$F(2, 1) = \int_{-\infty}^{2}\int_{-\infty}^{1} f(x, y)\, dy\, dx = \int_0^2\int_0^1 6e^{-2x-3y}\, dy\, dx = (1 - e^{-4})(1 - e^{-3}) = 0.9328.$$

c) Pre hustotu zložky X máme: Ak $x < 0$, tak $f_1(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_0^6 0\, dy = 0$.

Pre $x > 0$ platí $f_1(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_{-\infty}^{0} 0\, dy + \int_0^{\infty} e^{-2x-3y}\, dy = 0 + 2e^{-2x} = 2e^{-2x}$.

To znamená, že $X \sim \text{Exp}(\lambda)$, $\lambda = 2$. Analogicky dostaneme $Y \sim \text{Exp}(\lambda)$, $\lambda = 3$.

d) Podľa bodu (e) vety 3.1.9 máme: $P((X, Y) \in A) =$

$$= \iint_A f(x, y)\, dx\, dy = \int_{-\infty}^{\infty}\int_{-\infty}^{1-x} f(x, y)\, dx\, dy = \int_0^1\int_0^{1-x} 6e^{-2x-3y}\, dy\, dx = 1 - 3e^{-2} + 2e^{-3} = 0.6936$$

Vzhľadom na limitované ciele nášho textu, v prípade náhodného vektora spojitej povahy sa v ďalších článkoch budeme venovať len prípadu, keď zložky vektora sú nezávislé veličiny. Aj náš príklad 3.1.10 je situáciou, v ktorej zložky sú nezávislé veličiny.

Na záver tohoto článku predstavíme určite najdôležitejší prípad dvojrozmerného vektora s hustotou, a to prípad vektora s normálnym rozdelením.

3.1.11 Definícia. Náhodný vektor (X, Y) s hustotou má dvojrozmerné normálne rozdelenie, ak pre všetky $x, y \in R$ je jeho hustota daná vzťahom

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\, exp\left(\frac{-1}{2(1-\rho^2)}\left[\left(\frac{x-a}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-a}{\sigma_1}\right)\left(\frac{y-b}{\sigma_2}\right) + \left(\frac{y-b}{\sigma_2}\right)^2\right]\right)$$

kde $a, b \in R$, $\sigma_1 > 0$, $\sigma_2 > 0$ a $\rho \in (-1, 1)$ sú parametre. Symbolicky: $(X, Y) \sim N(a, b, \sigma_1^2, \sigma_2^2, \rho)$. Dá sa ukázať (podľa 3.1.9 b), že ak $(X, Y) \sim N(a, b, \sigma_1^2, \sigma_2^2, \rho)$, tak pre rozdelenia zložiek platí

$$X \sim N(a, \sigma_1^2),\ Y \sim N(b, \sigma_2^2).$$

Zložky X, Y môžu, ale aj nemusia, byť nezávislé. Informácia o tom je skrytá v parametri ρ. O jeho úlohe, resp. význame, sa dozvieme v článku 3.2 a v kapitole 4.

Úlohy

3.1.1 Nech pravdepodobnostná funkcia náhodného vektora (X, Y) je daná tabuľkou

$X \backslash Y$	-2	-1	0	1	2
-1	0.1	0.1	0.05	0	0
0	0	a	0.1	0.1	0.05
1	0	0	0.05	0.1	b

a) Určte a, b tak, aby rozdelenie zložky X bolo symetrické, t. j. aby $P(X=-1) = P(X=1)$.
b) Určte $P(X < Y)$, $P(X = Y)$, $P(X > Y)$ a $P(X + Y > 0)$.

3.1.2 Nasledujúce tabuľky určujú rozdelenia vektorov (X, Y), (U, V), (S, T). Všimnime si, že nielen hodnotové množiny prvých zložiek (veličín X, U, S) sú rovnaké, ale aj rozdelenia týchto veličín sú rovnaké. To isté platí o druhých zložkách! Samozrejme, ide o tri náhodné vektory, ktoré majú rôzne rozdelenia (hoci marginálne rozdelenia sú identické).

(X, Y):

	1	4	9
1	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
2	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
3	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$

(U, V):

	1	4	9
1	$\frac{2}{9}$	$\frac{1}{9}$	0
2	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
3	0	$\frac{1}{9}$	$\frac{2}{9}$

(S, T):

	1	4	9
1	$\frac{3}{9}$	0	0
2	0	$\frac{3}{9}$	0
3	0	0	$\frac{3}{9}$

Nájdite ďalší náhodný vektor, ktorého rozdelenie bude rôzne od rozdelení týchto troch vektorov, a pritom jeho zložky budú mať rovnaké rozdelenia ako majú zložky uvedených vektorov.

3.1.3 Sú dané tabuľky pravdepodobnostných funkcií $f_1(.)$, $f_2(.)$. Nájdite tabuľky dvoch náhodných vektorov, ktorých marginálne rozdelenia budú zhodné s danými $f_1(.)$, $f_2(.)$.

x_i	0	1	2
$f_1(x_i)$	0.3	0.4	0.3

y_j	0	1	2	4
$f_2(y_j)$	0.2	0.2	0.4	0.2

3.1.4 Náhodný pokus spočíva v hode dvoma normálnymi hracími kockami. Nech X znamená minimum bodov, ktoré padli a Y nech je súčet bodov.
a) Nájdite pravdepodobnostné funkcie veličín X, Y.
b) Nájdite pravdepodobnostnú funkciu vektora (X, Y). Presvedčte sa, že z nej získané marginálne funkcie sú tie, ktoré ste našli v bode (a).

3.1.5 Náhodný pokus spočíva v hode dvoma normálnymi hracími kockami. Nech X znamená absolútnu hodnotu rozdielu bodov, ktoré padli a Y nech je ich súčet.
a) Nájdite pravdepodobnostné funkcie veličín X, Y.
b) Nájdite pravdepodobnostnú funkciu vektora (X, Y). Presvedčte sa, že z nej získané marginálne funkcie sú tie, ktoré ste získali v bode (a).

3.1.6 Rub mince označme nulou a líc jednotkou. Hoďme mincou štyrikrát a s každým hodom spojme predstavu veličiny X_i, $H(X_i) = \{0, 1\}$, podľa toho, či v i-tom hode padne rub, alebo líc. Veličiny X_i sú navzájom rôzne, dokonca (intuitívne cítime) sú nezávislé (o nezávislosti hovoríme v nasledujúcom článku). Nech $U = X_1 + X_2$ a $V = X_3 + X_4$.
a) Je rozdelenie veličín U a V rovnaké? Nájdite ich pravdepodobnostné funkcie.
b) Nájdite pravdepodobnostnú funkciu vektora (U, V) a presvedčte sa, že z nej získané marginálne funkcie sú tie, ktoré ste dostali v bode (a).

3.1.7 Rub mince označme nulou a líc jednotkou. Hoďme mincou štyrikrát a s každým hodom spojme predstavu veličiny X_i tak, ako v predchádzajúcej úlohe. Veličiny U a V definujeme teraz inak: $U = X_1 + X_2 + X_3$ a $V = X_2 + X_3 + X_4$.
a) Je rozdelenie veličín U a V rovnaké? Nájdite ich pravdepodobnostné funkcie.
b) Nájdite pravdepodobnostnú funkciu vektora (U, V) a presvedčte sa, že z nej získané marginálne funkcie sú tie, ktoré ste dostali v bode (a).

3.1.8 V škatuli máme 7 bielych a 3 čierne loptičky. Náhodne vytiahneme naraz tri. Nech X je počet bielych (medzi vytiahnutými) a Y je počet čiernych (medzi vytiahnutými).
a) Nájdite pravdepodobnostné funkcie veličín X a Y.
b) Nájdite pravdepodobnostnú funkciu vektora (X, Y).

3.1.9 Funkcia f je daná vzťahom

$$f(x, y) = \begin{cases} x + y, & x \in \langle 0, 1 \rangle, y \in \langle 0, 1 \rangle, \\ 0, & \textit{inde}. \end{cases}$$

a) Nájdite hustoty zložiek vektora.
b) Vypočítajte $P(0.5 \le X < 1, 0 \le Y < 0.5)$.
c) Vypočítajte $P(X < Y)$.

3.1.10 Funkcia f je daná vzťahom

$$f(x, y) = \begin{cases} k(4x - y), & x \in \langle 1, 2 \rangle, y \in \langle 2, 4 \rangle, \\ 0, & \textit{inde}. \end{cases}$$

a) Určte hodnotu konštanty tak, aby f bola hustota náhodného vektora.
b) Nájdite hustoty oboch zložiek vektora.
c) Vypočítajte $P(1 \le X < 1.5, 3 \le Y < 4)$.
d) Vypočítajte $P(2X < Y)$.

3.2 Nezávislosť náhodných veličín

V kapitole 1 sme definovali nezávislosť náhodných udalostí. Vieme, že udalosti A, B sú nezávislé (t. j. stochasticky nezávislé) práve vtedy, ak platí: $P(A \cap B) = P(A)P(B)$. Ak X, Y sú náhodné veličiny (na tom istom pravdepodobnostnom priestore), tak je prirodzené považovať ich za *nezávislé* práve vtedy, keď udalosti vyjadrené prostredníctvom veličiny X, t. j. udalosti typu $\{a \le X < b\}$, sú nezávislé s udalosťami typu $\{c \le Y < d\}$, t. j. s udalosťami, ktoré vypovedajú len o veličine Y. Takáto skutočnosť je obsahom nasledujúcej definície.

3.2.1 Definícia. Veličiny X, Y (definované na tom istom pravdepodobnostnom priestore) sú nezávislé práve vtedy, keď pre všetky $a, b, c, d \in R$ platí

$$P(a \le X < b, c \le Y < d) = P(a \le X < b) P(c \le Y < d).$$

Definícia je v zhode s intuitíciou, ale uvážme, že rozdelenia, t. j. systémy pravdepodobností $P(a \le X < b)$, resp. $P(c \le Y < d)$, resp. $P(a \le X < b, c \le Y < d)$ sú známe prostredníctvom funkcií – a to buď distribučných funkcií, alebo pravdepodobnostných funkcií (resp. hustôt). Preto je potrebné formulovať podmienku nezávislosti veličín X, Y pomocou týchto funkcií.

3.2.2 Veta. Nech $F(.,.)$ je distribučná funkcia vektora (X, Y) a F_1, F_2 sú distribučné funkcie zložiek. Veličiny X, Y sú nezávislé práve vtedy, keď platí

$$F(x,y) = F_1(x)\,F_2(y), \quad \text{pre všetky } x, y \in \mathrm{R}.$$

Dôkaz. Namiesto úplného dôkazu ukážme len to, že podmienka $F(x,y) = F_1(x)F_2(y)$, pre všetky $x, y \in \mathrm{R}$, je postačujúca (pre nezávislosť X a Y). Najprv využijeme poznámku 3.1.2 a v druhom kroku predpoklad: $F(x,y) = F_1(x)F_2(y)$, pre všetky $x, y \in \mathrm{R}$. Postupne máme

$$\begin{aligned} P(a \le X < b, c \le Y < d) &= F(b,d) - F(a,d) - F(b,c) + F(a,c) = \\ &= F_1(b)F_2(d) - F_1(a)F_2(d) - F_1(b)F_2(c) + F_1(a)F_2(c) = \\ &= F_2(d)(F_1(b) - F_1(a)) - F_2(c)(F_1(b) - F_1(a)) = \\ &= (F_1(b) - F_1(a))(F_2(d) - F_2(c)) = P(a \le X < b)\,P(c \le Y < d). \end{aligned}$$

Nasledujúca veta formuluje nezávislosť pomocou pravdepodobnostných funkcií (resp. pomocou hustôt).

3.2.3 Veta. (a) Nech (X, Y) je diskrétny vektor a f_1, f_2 sú pravdepodobnostné funkcie zložiek. Veličiny X, Y sú nezávislé práve vtedy, ak funkcia g definovaná vzťahom

$$g(x,y) = f_1(x)\,f_2(y), \quad \text{pre všetky } x, y \in \mathrm{R},$$

je pravdepodobnostnou funkciou vektora (X, Y).

(b) Nech (X, Y) je vektor s hustotou a f_1, f_2 sú hustoty zložiek. Zložky X, Y sú nezávislé práve vtedy, ak funkcia g definovaná vzťahom

$$g(x,y) = f_1(x)\,f_2(y), \quad \text{pre všetky } x, y \in \mathrm{R},$$

je hustotou vektora (X, Y).

Dôkaz časti (a). Nech X, Y sú nezávislé podľa definície 3.2.1. Máme ukázať, že napr. platí $P(X = x_1, Y = y_1) = f_1(x_1)f_2(y_1)$. V definícii 3.2.1 položme $a = x_1$, $b = x_2$, $c = y_1$, $d = y_2$. Zrejme

$$P(X = x_1, Y = y_1) = P(x_1 \le X < x_2, y_1 \le Y < y_2) = P(x_1 \le X < x_2)\,P(y_1 \le Y < y_2) = f_1(x_1)\,f_2(y_1).$$

Analogickými voľbami ľahko získame platnosť všetkých požadovaných rovností

$$P(X = x_i, Y = y_j) = f_1(x_i)\,f_2(y_j), \quad \text{pre } x_i \in H(X),\ y_j \in H(Y).$$

A teraz naopak, nech platia rovnosti $P(X = x_i, Y = y_j) = f_1(x_i)f_2(y_j)$, pre $x_i \in H(X)$, $y_j \in H(Y)$. Treba ukázať, že platí: $P(a \le X < b, c \le Y < d) = P(a \le X < b)\,P(c \le Y < d)$, pre všetky $a, b, c, d \in \mathrm{R}$. Pre jednoduchosť, nech z hodnôt veličiny X v intervale $\langle a, b)$ ležia, povedzme, body x_2, x_3 a x_4. Analogicky, z hodnôt veličiny Y, nech v intervale $\langle c, d)$ ležia body y_3, y_4, y_5, y_6. Zrejme

$$P(a \le X < b, c \le Y < d) = \textstyle\sum_{i=2}^{4}\sum_{j=3}^{6} f(x_i, y_j) = \sum_{i=2}^{4}\sum_{j=3}^{6} f_1(x_i) f_2(y_j) = \sum_{i=2}^{4} f_1(x_i) \sum_{j=3}^{6} f_2(y_j) =$$

$$= \textstyle\sum_{i=2}^{4} f_1(x_i)\, P(c \le Y < d) = P(c \le Y < d) \sum_{i=2}^{4} f_1(x_i) = P(c \le Y < d)\, P(a \le X < b).$$

Tým je dôkaz časti (a) skončený. Aj keď prípad diskrétneho vektora (zvlášť ak obe zložky majú konečne veľa hodnôt) je vskutku jednoduchý, je dobré uvedomiť si, že ak máme pred sebou vektor (X, Y), tak môžeme byť v dvoch polohách. Alebo

- máme dané rozdelenie vektora, t. j. je daná tabuľka funkcie $f(.,.)$ – v takom prípade otázku o tom, či zložky sú nezávislé, zodpovieme jednoznačne po overení rovností
$$f(x_i, y_j) = f_1(x_i)\,f_2(y_j), \quad \text{pre všetky } x_i \in H(X),\ y_j \in H(Y), \text{ alebo}$$
- rozdelenie vektora známe nie je, sú však známe rozdelenia zložiek, teda tabuľky funkcií $f_1(.)$, $f_2(.)$ a ako *podmienka* je dané, že zložky X, Y sú nezávislé. V takom prípade za pravdepodobnostnú funkciu vektora vezmeme funkciu g definovanú takto:
$$g(x_i, y_j) = f_1(x_i)\,f_2(y_j), \quad \text{pre } x_i \in H(X),\ y_j \in H(Y).$$

3.2.4 Príklad. Hore uvedené ilustrujme jednoduchou situáciou. Tabuľky predstavujú rozdelenia dvoch náhodných vektorov (X, Y) a (U, V). V ktorom z nich sú zložky vektora nezávislé?

X/Y	0	1	2
0	0.05	0.10	0.10
1	0.15	0.30	0.30

U/V	0	1	2
0	0.05	0.30	0.10
1	0.15	0.10	0.30

Ľahko sa overí, že kým zložky vektora (X, Y) sú nezávislé, zložky vektora (U, V) sú závislé. A teraz situáciu, ktorá ilustruje druhú polohu. Máme dané jednorozmerné tabuľky, ktoré predstavujú rozdelenia *zložiek* vektora (S, T), o ktorom vieme, že jeho zložky S, T sú nezávislé veličiny. Nájdite dvojrozmernú tabuľku, ktorá predstavuje rozdelenie vektora (S, T).

s_i	0	1
$f_1(s_i)$	0.4	0.6

t_j	0	1	2	3
$f_2(t_j)$	0.2	0.2	0.3	0.3

(Táto úloha je obsahom úlohy 3.2.3.)

3.2.5 Poznámka. Obsah Vety 3.2.3 často uplatňujeme pri modelovaní reálnych situácií, v ktorých nezávislosť náhodných veličín, povedzme, X a Y, vyplýva z pozadia experimentu. Vzhľadom na Vetu 3.2.3, správne modelovanie takej dvojice X, Y spočíva v tom, že rozdelenie vektora (X, Y) definujeme ako súčin marginálnych rozdelení. Pritom predstavu o marginálnych rozdeleniach, teda o rozdeleniach X a Y, získavame relatívne ľahko. Toto ilustrujú príklady 3.2.6, 3.2.10 a dopĺňa poznámka 3.2.9.

3.2.6 Príklad. (Ilustrácia poznámky 3.2.5) Predpokladajme, že experimenty potvrdzujú, že životnosť žiariviek istého typu môžeme modelovať rozdelením $\mathrm{Exp}(\lambda)$, kde $\lambda = 0.01$. Náhodne sme vybrali dve žiarivky a chceme

a) vyčísliť pravdepodobnosť toho, že každá z nich bude pracovať aspoň 150 hodín.
b) Ak druhú žiarivku použijeme v prístroji po zlyhaní (zničení) prvej, aká je pravdepodobnosť toho, že prístroj bude mať fungujúcu žiarivku aspoň 300 hodín?

Riešenie. Nech veličiny X, resp. Y modelujú životnosti žiariviek. Pretože $X, Y \sim \mathrm{Exp}(0.01)$, vieme, že

$$P(X > 150) = 1 - P(X < 150) = 1 - (1 - \exp(-0.01 \cdot 150)) = \exp(-1.5) = 0.22313.$$

To isté platí zrejme aj pre veličinu Y. Avšak X a Y považujeme za nezávislé (žiarivky sme vybrali náhodne z celkovej produkcie), preto udalosti $\{X > 150\}$, $\{Y > 150\}$ sú nezávislé, a tak

$$\mathrm{P}(X > 150,\ Y > 150) = \mathrm{P}(X > 150)\ \mathrm{P}(Y > 150) = 0.22313^2 = 0.04979.$$

Časť (b). Teraz ide o $\mathrm{P}(X + Y > 300)$. Kým prvá otázka mohla byť položená v kontexte, keď máme dva prístroje (v každom jednu žiarivku) a pýtame sa, s akou pravdepodobnosťou bude v každom z prístrojov žiarivka fungovať aspoň 150 hodín. Teraz ide o situáciu, v ktorej máme jeden prístroj s pôvodnou a jednou náhradnou žiarivkou a pýtame sa, s akou pravdepodobnosťou sa počas 300 hodín nedostaneme do ťažkostí (týkajúcich sa vypálenej žiarivky). Odpoveď nájdeme v článku 3.3 (dozvieme sa, aké rozdelenie má veličina $X + Y$).

3.2.7 Je prirodzené pýtať sa, či z nezávislosti veličín X a Y vyplýva nezávislosť napr. veličín $2X - 1$ a $3Y + 5$. Nie je ťažké ukázať to priamo z definície 3.2.1. Dokonca, rovnako by bolo možné ukázať, že z nezávislosti X a Y vyplýva nezávislosť veličín $aX + b$ a $cY + d$, kde a, b, c, d sú ľubovoľné čísla. Ide o intuitívne akceptovateľné skutočnosti, ale namiesto formálneho dôkazu sformulujeme vetu, pomocou ktorej rozhodneme aj oveľa náročnejšie situácie.

3.2.8 Veta. Nech X, Y sú náhodné veličiny a h_1, h_2: $\mathrm{R} \to \mathrm{R}$ sú také, že $h_1(X)$ a $h_2(Y)$ sú náhodné veličiny. Ak X, Y sú nezávislé, potom sú nezávislé aj veličiny $h_1(X)$ a $h_2(Y)$.

Dôkaz vety je nad naše sily, ale jej obsah je jasný. Ak X a Y sú nezávislé, tak nezávislými sú aj X^2 a Y^2, alebo napr. $\sin(X)$ a $\sin(Y)$ (veta nehovorí, že h_1, h_2 musia byť rôzne), alebo X^2 a $\sqrt{Y}$, alebo napr. $\sin(X)$ a $\cos(Y)$, a pod.

3.2.9 Poznámka. (a) Definície, resp. vety odstavcov 3.2.1 až 3.2.8 majú prirodzené zovšeobecnenia pre viacrozmerné vektory. Napr. X_1, X_2, X_3 sú nezávislé práve vtedy, ak platí

$$\mathrm{P}(a_1 \le X_1 < b_1,\ a_2 \le X_2 < b_2,\ a_3 \le X_3 < b_3) = \mathrm{P}(a_1 \le X_1 < b_1)\,\mathrm{P}(a_2 \le X_2 < b_2)\,\mathrm{P}(a_3 \le X_3 < b_3)$$

pre všetky $a_1, b_1, a_2, b_2, a_3, b_3 \in \mathrm{R}$. Vďaka spojitosti a aditivite pravdepodobnosti uvedená podmienka je ekvivalentná každej z nasledujúcich podmienok:

Pre všetky $c_1, c_2, c_3 \in \mathrm{R}$ platí: $\mathrm{P}(X_1 < c_1, X_2 < c_2, X_3 < c_3) = \mathrm{P}(X_1 < c_1)\,\mathrm{P}(X_2 < c_2)\,\mathrm{P}(X_3 < c_3)$.

Pre všetky $c_1, c_2, c_3 \in \mathrm{R}$ platí: $\mathrm{P}(X_1 > c_1, X_2 > c_2, X_3 > c_3) = \mathrm{P}(X_1 > c_1)\,\mathrm{P}(X_2 > c_2)\,\mathrm{P}(X_3 > c_3)$.

(b) Ďalej napr. ak $(X_1, X_2, \dots, X_n)$ je náhodný vektor a ak f_i je hustota X_i, tak $X_1, X_2, \dots, X_n$ sú nezávislé práve vtedy, keď funkcia

$$g(x_1, x_2, \dots, x_n) = f_1(x_1) f_2(x_2) \dots f_n(x_n)$$

je hustota vektora $(X_1, X_2, \dots, X_n)$. Úplne rovnako to platí aj v prípade, keď ide o diskrétny náhodný vektor, ktorého rozdelenie určuje pravdepodobnostná funkcia.

V štatistike je časté modelovanie situácie, v ktorej X_i sú nezávislé náhodné veličiny s *rovnakým* rozdelením, napr. s rovnakou hustotou f. Pritom nezávislosť X_i sa predpokladá, pretože je intuitívne zrejmá z pozadia konkrétneho štatistického experimentu. Modelovanie takej situácie pozostáva potom z toho, že za hustotu vektora $(X_1, X_2, \dots, X_n)$ berieme funkciu

$$g(x_1, x_2, \dots, x_n) = f(x_1) f(x_2) \dots f(x_n)$$

(na pravej strane je súčin), resp. pre distribučnú funkciu G vektora $(X_1, X_2, \dots, X_n)$ kladieme

$$G(x_1, x_2, \dots, x_n) = F(x_1) F(x_2) \dots F(x_n),$$

kde F je distribučná funkcia X_i, pre $i = 1, 2, \dots, n$ (opäť na pravej strane rovnosti je súčin).

3.2.10 Príklad. Náhodne vyberieme dva výrobky z produkcie výrobkov, ktorých životnosť modeluje rozdelenie N(210, 64). Potom náhodne vyberieme ďalšie dva, tentoraz z produkcie takých, ktorých životnosť modeluje rozdelenie N(218, 81). Aká je pravdepodobnosť toho, že každý vybratý výrobok bude fungovať aspoň 200 hodín?

Riešenie. Ak životnosť prvých dvoch vybratých modelujú veličiny X_1, X_2, potom

$$\mathrm{P}(X_1 > 200) = 1 - \mathrm{P}(X_1 < 200) = 1 - F_N\left(\frac{200 - 210}{8}\right) = 1 - F_N(-1.25) = F_N(1.25) = 0.8944.$$

To isté platí pre $\mathrm{P}(X_2 > 200)$. Analogicky, nech X_3, X_4, modelujú životnosti ďalších dvoch vybratých. Máme

$$\mathrm{P}(X_3 > 200) = 1 - \mathrm{P}(X_3 < 200) = 1 - F_N\left(\frac{200 - 218}{9}\right) = 1 - F_N(-2) = F_N(2) = 0.9773,$$

pričom rovnaký výsledok platí pre $\mathrm{P}(X_4 > 200)$. Nakoniec, vďaka nezávislosti, pozri časť a) poznámky 3.2.9, máme

$$\mathrm{P}(X_1 > 200, X_2 > 200, X_3 > 200, X_4 > 200) = (0.8944)^2\,(0.9773)^2 = 0.7640.$$

3.2.11 Veta. Nech $(X, Y) \sim \mathrm{N}(a,\ b,\ \sigma_1^2,\ \sigma_2^2,\ \rho)$, t. j. (X, Y) má dvojrozmerné normálne rozdelenie s hustotou (def. 3.1.11). Zložky X, Y sú nezávislé práve vtedy, keď parameter $\rho = 0$.

Dôkaz. Nech $(X, Y) \sim \mathrm{N}(a, b, \sigma_1^2, \sigma_2^2, \rho)$. Vieme (podľa 3.1.11), že pre rozdelenia zložiek platí $X \sim \mathrm{N}(a, \sigma_1^2)$, $Y \sim \mathrm{N}(b, \sigma_2^2)$. Teraz predpokladajme, že X, Y sú nezávislé (a chceme dokázať, že $\rho = 0$). Podľa vety 3.2.3 súčin hustôt zložiek je hustotou vektora, čo znamená, že pre hustotu vektora (X, Y) dostávame

$$f(x, y) = f_1(x)\, f_2(y) = \frac{1}{\sqrt{2\pi}\,\sigma_1}\, e^{-\frac{(x-a)^2}{2\sigma_1^2}}\ \frac{1}{\sqrt{2\pi}\,\sigma_2}\, e^{-\frac{(y-b)^2}{2\sigma_2^2}} = \frac{1}{2\pi\,\sigma_1\sigma_2}\, e^{-\frac{(x-a)^2}{2\sigma_1^2} - \frac{(y-b)^2}{2\sigma_2^2}}$$

Z tvaru získanej funkcie je zrejmé, že ide o hustotu rozdelenia $\mathrm{N}(a, b, \sigma_1^2, \sigma_2^2, \rho)$, v ktorom pre parameter ρ platí: $\rho = 0$.

Dôkaz opačnej implikácie je ešte jednoduchší. Ak $(X, Y) \sim \mathrm{N}(a, b, \sigma_1^2, \sigma_2^2, \rho)$ a $\rho = 0$, tak hustota vektora je evidentne rovná súčinu marginálnych hustôt, a preto podľa vety 3.2.3 sú X a Y nezávislé.

Úlohy

3.2.1 Rozhodnite, či zložky vektora (X, Y) sú nezávislé, keď jeho rozdelenie je dané tabuľkou

$X \backslash Y$	-1	0	1
-1	.08	.08	.04
0	.20	.20	.10
1	.12	.13	.05

3.2.2 Je daná tabuľka

$X\backslash Y$	0	2	4	6
1	.04	.08	a	.16
3	.06	b	.18	.24

a) Je možné stanoviť a, b tak, aby tabuľka bola tabuľkou pravdepodobnostnej funkcie náhodného vektora? Je takých možností viac?

b) Je možné stanoviť a, b tak, aby X, Y boli nezávislé? Je takých možností viac?

3.2.3 Sú dané jednorozmerné tabuľky, ktoré predstavujú rozdelenia zložiek vektora (S, T). Nájdite tabuľku rozdelenia vektora (S, T), ak vieme, že zložky S, T sú nezávislé veličiny.

s_i	0	1
$f_1(s_i)$	0.4	0.6

t_j	0	1	2	3
$f_2(t_j)$	0.2	0.2	0.3	0.3

3.2.4 Nájdite pravdepodobnostnú funkciu vektora (X, Y) tak, aby zložky vektora boli nezávislé, a pritom pre ich rozdelenia má platiť:

a) $X \sim R\{0, 1, 2, 3\}$, $Y \sim R\{1, 2, 3\}$
b) $X \sim Bi(3, 0.5)$, $Y \sim R\{-1, 1\}$
c) $X \sim Bi(3, 0.3)$, $Y \sim Bi(2, 0.3)$
d) $X \sim Po(2)$, $Y \sim Po(3)$
e) $X \sim G(0.6)$, $Y \sim A(0.3)$
f) $X \sim G(p_1)$, $Y \sim G(p_2)$

3.2.5 Nech $X_1, X_2, \dots, X_n$ sú nezávislé veličiny. Modifikujte vetu 3.2.8 tak, aby sme sa mohli na ňu odvolať a tvrdiť, že potom aj veličiny $X_1^2, X_2^2, \dots, X_n^2$ sú nezávislé.

3.2.6 Náhodný vektor (X, Y) má hustotu f danú vzťahom

$$f(x,y) = \begin{cases} \frac{1}{4}(1 + xy(x^2 - y^2)), & x, y \in (-1, 1), \\ 0, & inak. \end{cases}$$

Nájdite hustoty zložiek a rozhodnite, či X, Y sú nezávislé.

3.2.7 Nájdite funkciu hustoty vektora (X, Y) tak, aby zložky vektora boli nezávislé a pre ich rozdelenia má platiť:

a) $X \sim R(-1, 1)$, $Y \sim R(-1, 1)$
b) $X \sim Exp(\lambda)$, $Y \sim Exp(\lambda)$
c) $X \sim N(0, 1)$, $Y \sim N(0, 1)$
d) $X \sim N(1, 4)$, $Y \sim N(-2, 9)$

Vyčíslite $P(a \le X < b, c \le Y < d)$, pre vhodné $a, b, c, d \in R$.

3.2.8 Predpokladajme, že náhodné chyby sprevádzajúce meranie vzdialenosti majú rozdelenie $N(0, \sigma^2)$. Odmerajme tri razy tú istú vzdialenosť d tak, aby merania, resp. chyby pri meraniach mohli byť považované za nezávislé. S akou pravdepodobnosťou

a) prvé a druhé meranie dajú výsledky menšie ako $d + \sigma$ a súčasne tretie meranie výsledok menší ako $d + 2\sigma$?

b) každá z chýb sprevádzajúca meranie bude v absolútnej hodnote menšia ako 2σ?

3.3 Transformácia náhodného vektora

Ak X, Y sú náhodné veličiny na (Ω, S, P), potom náhodnými veličinami sú aj $Z = X + Y$, resp. $U = XY$, $V = X^2 + Y^2$, alebo $W = min(X, Y)$. Aj keď intuitívne je toto konštatovanie zrejmé, dôkaz nie je jednoduchý (leží však mimo nášho záujmu). Tieto veličiny vznikli transformáciou vektora (X, Y) prostredníctvom funkcií

$$h_1(x,y) = x + y,\ h_2(x,y) = xy,\ h_3(x,y) = x^2 + y^2,\ h_4(x,y) = min(x,y).$$

Cieľ tohoto článku je odvodiť rozdelenie veličiny $Z = h(X, Y)$, keď poznáme rozdelenie vektora (X, Y) a samozrejme, je daná funkcia $h(.,.)$, ktorá realizuje transformáciu. Najprv preberieme konkrétne situácie a potom sformulujeme všeobecný postup. Začneme prípadom diskrétneho vektora, keď úloha je relatívne jednoduchá. Kľúčovou vlastnosťou pri odvádzaní rozdelenia diskrétnej veličiny $Z = h(X, Y)$ je aditivita pravdepodobnosti.

3.3.1 Príklad. Pravdepodobnostná funkcia vektora (X, Y) je daná tabuľkou

$X\backslash Y$	0	1	2
-1	0.10	0.05	0.05
0	0.15	0	0.15
1	0.20	0.20	0.10

Nájdime rozdelenie náhodných veličín $Z = X + Y$, $U = XY$, $V = \sqrt{X^2 + Y^2}$, $W = min(X, Y)$.

Riešenie. Pretože tabuľka je malých rozmerov, nie je potrebné nič algoritmizovať a napr.

$$P(X + Y = 1) = P(\{(X,Y) \in \{(-1, 2), (0, 1), (1, 0)\}\}) = 0.05 + 0 + 0.20 = 0.25.$$

Takto získame rozdelenie $Z = X + Y$. Je dané tabuľkou

z	-1	0	1	2	3
$P(Z = z)$	0.10	0.20	0.25	0.35	0.10

Analogicky nájdeme rozdelenie $U = XY$. Napríklad $P(U = 0)$ stanovíme takto:

$P(XY = 0) = P(\{(X,Y) \in \{(-1, 0), (0, 0), (0, 1), (0, 2), (1, 0)\}\}) = 0.10 + 0.15 + 0 + 0.15 + 0.20 = 0.60.$

Rozdelenie veličiny $U = XY$ určuje tabuľka

u	-2	-1	0	1	2
$P(U = u)$	0.05	0.05	0.60	0.20	0.10

Overte, že rozdelenie $V = \sqrt{X^2 + Y^2}$ je dané tabuľkou

v	0	1	$\sqrt{2}$	2	$\sqrt{5}$
$P(V = v)$	0.15	0.30	0.25	0.15	0.15

Nakoniec, pravdepodobnostnú funkciu veličiny $W = min(X, Y)$ určuje tabuľka

w	-1	0	1
$P(W = w)$	0.2	0.5	0.3

Napr. $P(W = 0) = P((X,Y) \in \{(0, 0), (0, 1), (0, 2), (1, 0)\}) = 0.15 + 0 + 0.15 + 0.20 = 0.50.$

3.3.2 Odvoďme vzťah pre pravdepodobnostnú funkciu g veličiny $Z = h(X, Y)$, v prípade diskrétneho vektora (X, Y) s pravdepodobnostnou funkciou $f(.,.)$.

$$g(z) = \mathrm{P}(Z = z) = \mathrm{P}(\,h(X, Y) = z\,) = \textstyle\sum_{(x,y)\in A_z} f(x, y), \quad \text{kde } A_z = \{\,(x, y)\colon h(x, y) = z\,\}.$$

Základnú transformáciu predstavuje súčet, $Z = X + Y$. V takom prípade

$$g(z) = \mathrm{P}(\,X + Y = z\,) = \textstyle\sum_{(x,y)\in A_z} f(x, y) = \sum_{(x,y):\, x+y=z} f(x, y) = \sum_{x\in H(X)} f(x, z - x),$$

pretože v prípade súčtu, $h(x, y) = x + y$, máme $A_z = \{\,(x, y) : h(x, y) = z\,\} = \{\,(x, y) : x + y = z\,\}$.

3.3.3 Príklad. Nech pravdepodobnostná funkcia $f(.,.)$ vektora (X, Y) je daná tabuľkou

$X\backslash Y$	0	1	2	3
1	0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
2	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{2}{12}$	$\frac{1}{12}$
3	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	0

Určíme rozdelenie veličiny $Z = X + Y$.

Riešenie. Nech g je pravdepodobnostná funkcia Z. Napríklad, pre $g(3)$ podľa 3.3.2 platí

$$g(3) = \textstyle\sum_{(x,y):\, x+y=3} f(x, y) = f(1, 2) + f(2, 1) + f(3, 0) = \frac{1}{12} + \frac{2}{12} + \frac{1}{12} = \frac{4}{12}.$$

Takto získame tabuľku funkcie g

z	2	3	4	5
$g(z)$	$\frac{2}{12}$	$\frac{4}{12}$	$\frac{4}{12}$	$\frac{2}{12}$

3.3.4 Poznámka. Dôležitým prípadom je situácia, keď X a Y sú nezávislé. Vtedy pravdepodobnostná funkcia vektora $f(.,.)$ je súčinom marginálnych pravdepodobnostných funkcií f_1 a f_2, to znamená, že pre všetky $x, y \in \mathrm{R}$ platí $f(x, y) = f_1(x)\, f_2(y)$. Ak ide o rozdelenie súčtu $X + Y$, tak pre pravdepodobnostnú funkciu g veličiny $Z = X + Y$ máme

$$g(z) = \textstyle\sum_{x\in H(X)} f_1(x) f_2(z - x) = \sum_{y\in H(Y)} f_1(z - y) f_2(y),$$

pre všetky $z \in \mathrm{R}$.

3.3.5 Príklad. Nech pre nezávislé X, Y platí: $X \sim \mathrm{R}\{\,1, 2, 3\,\}$, $Y \sim \mathrm{R}\{\,0, 1, 2, 3\,\}$. Nájdime rozdelenie veličiny $Z = X + Y$.

Riešenie. Na rozdiel od príkladu 3.3.3, teraz nie je nutné pracovať s dvojrozmernou tabuľkou funkcie $f(.,.)$. Vystačíme s marginálnymi funkciami. Zrejme $f_1(x) = \frac{1}{3}$, pre $x \in \{1, 2, 3\}$ a $f_2(y) = \frac{1}{4}$, pre $y \in \{\,0, 1, 2, 3\,\}$. Podľa 3.3.4 napríklad, pre hodnotu $g(3)$ pravdepodobnostnej funkcie veličiny Z máme

$$g(3) = \textstyle\sum_{x\in H(X)} f_1(x) f_2(3 - x) = \sum_{x=1}^{3} f_1(x) f_2(3 - x) = \frac{1}{3}\frac{1}{4} + \frac{1}{3}\frac{1}{4} + \frac{1}{3}\frac{1}{4} = \frac{3}{12}.$$

Overte, že rozdelenie $X + Y$ určuje tabuľka

z	1	2	3	4	5	6
$g(z)$	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{3}{12}$	$\frac{3}{12}$	$\frac{2}{12}$	$\frac{1}{12}$

Všimnime si, že súčtom rovnomerných rozdelení, už nie je rovnomerné rozdelenie.

3.3.6 Príklad. Teraz uvažujeme súčty viacerých nezávislých veličín, pričom všetky majú rovnaké rozdelenie. Takáto situácia vzniká pri viacnásobnom (nezávislom) opakovaní toho istého pokusu (za rovnakých podmienok) a je kľúčová pre štatistické modelovanie.

Nech pokus spočíva v opakovanovaní hodu obyčajnou hracou kockou. Pokus s jedným hodom modeluje veličina X, $X \sim \mathrm{R}\{1, 2, \dots, 6\}$. Pri dvojnásobnom hode kockou, v ktorom nás zaujíma súčet padnutých bodov, výsledok modeluje veličina S_2

$S_2 = X_1 + X_2$, kde X_1 predstavuje počet bodov na kocke v prvom hode a
X_2 predstavuje počet bodov na kocke v druhom hode.

Veličiny X_1, X_2 sú nezávislé a rovnako rozdelené (veď hody sú intuitívne nezávislé a vykonané za rovnakých podmienok). Zrejme $X_1, X_2 \sim \mathrm{R}\{\,1, 2, \dots, 6\,\}$. Rozdelenie S_2 nájdeme podľa 3.3.4 (tak ako v príklade 3.3.5). Je ľahké overiť, že je dané tabuľkou (obr. 3-1b)

2	3	4	5	6	7	8	9	10	11	12
$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

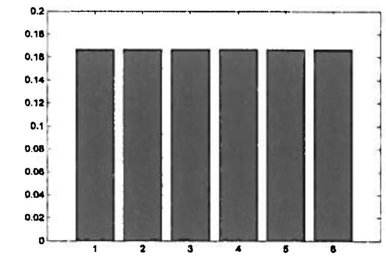
Keď pokus spočíva v tom, že kockou hodíme tri razy a S_3 predstavuje súčet bodov, tak

$S_3 = X_1 + X_2 + X_3$, kde X_i predstavuje počet bodov v i-tom hode.

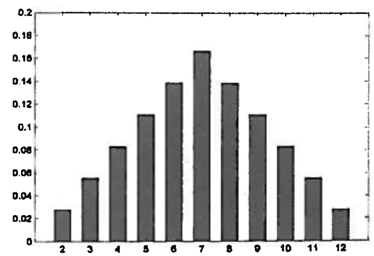
Rozdelenie S_3 určíme opäť podľa 3.3.4. Zrejme $S_3 = (X_1 + X_2) + X_3 = S_2 + X_3$, pravdepodobnostnú funkciu S_2 sme práve našli a $X_3 \sim \mathrm{R}\{\,1, 2, \dots, 6\,\}$. Rozdelenie S_3 určuje tabuľka

3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$\frac{1}{216}$	$\frac{3}{216}$	$\frac{6}{216}$	$\frac{10}{216}$	$\frac{15}{216}$	$\frac{21}{216}$	$\frac{25}{216}$	$\frac{27}{216}$	$\frac{27}{216}$	$\frac{25}{216}$	$\frac{21}{216}$	$\frac{15}{216}$	$\frac{10}{216}$	$\frac{6}{216}$	$\frac{3}{216}$	$\frac{1}{216}$

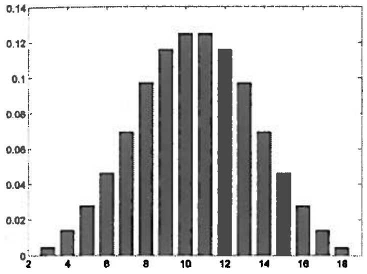
Rozdelenie X a všetkých X_i je rovnomerné. Rozdelenie $S_2 = X_1 + X_2$ má trojuholníkový tvar (obr. 3-1b). Rozdelenie $S_3 = X_1 + X_2 + X_3$ má už zvonový tvar (obr. 3-1c). Takýto tvar ešte lepšie ilustruje rozdelenie $S_4 = X_1 + X_2 + X_3 + X_4$, t. j. súčet v pokuse so štvornásobným opakovaním hodu (pravdepodobnostnú funkciu S_4 zobrazuje obr. 3-1d).



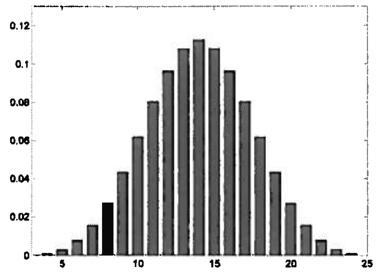
Obr. 3-1a. Rozdelenie veličiny X



Obr. 3-1b. Rozdelenie veličiny $X_1 + X_2$



Obr. 3-1c. Rozdelenie $X_1 + X_2 + X_3$



Obr. 3-1d. Rozdelenie $X_1 + X_2 + X_3 + X_4$

Vidíme, že súčty rovnomerne rozdelených veličín už nie sú veličiny s rovnomerným rozdelením. Avšak pre niektoré rozdelenia (za predpokladu nezávislosti) typ rozdelenia súčtu ostáva rovnaký (ako majú sčítance) a zmenia sa len parametre (väčšinou celkom prirodzeným a očakávaným spôsobom). Základné fakty formuluje veta 3.3.7.

3.3.7 Veta. Nech X, Y sú nezávislé náhodné veličiny.

1. Ak $X \sim \mathrm{Bi}(n, p)$, $Y \sim \mathrm{Bi}(m, p)$, tak $X + Y \sim \mathrm{Bi}(n + m, p)$.
2. Ak $X \sim \mathrm{Po}(\lambda_1)$, $Y \sim \mathrm{Po}(\lambda_2)$, tak $X + Y \sim \mathrm{Po}(\lambda_1 + \lambda_2)$.
3. Ak $X \sim \mathrm{Exp}(\lambda)$, $Y \sim \mathrm{Exp}(\lambda)$, tak $X + Y \sim \mathrm{Erl}(2, \lambda)$.
4. Ak $X \sim \mathrm{Erl}(n, \lambda)$, $Y \sim \mathrm{Erl}(m, \lambda)$, tak $X + Y \sim \mathrm{Erl}(n+m, \lambda)$.
5. Ak $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$, tak $X + Y \sim \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Dôkazy tvrdení z bodov 1, 2 možno urobiť použitím výsledku z odseku 3.3.4. Všimnime si, že rozdelenie $\mathrm{A}(p) = \mathrm{Bi}(1, p)$, a preto podľa bodu 1 pre nezávislé veličiny X, Y máme:

Ak $X \sim \mathrm{A}(p)$, $Y \sim \mathrm{A}(p)$, tak $X + Y \sim \mathrm{Bi}(2, p)$, čo môžeme zovšeobecniť na

1* Ak $X_i \sim \mathrm{A}(p)$, pre $i = 1, \dots, n$ a sú nezávislé, tak $\sum_{i=1}^{n} X_i \sim \mathrm{Bi}(n, p)$.

Dôkazy tvrdení z bodov 3, 4 a 5 sú mimo našich možností. Avšak budeme ich v ďalšom potrebovať, naviac, v takýchto všeobecnejších formuláciách:

3* Ak $X_i \sim \mathrm{Exp}(\lambda)$, pre $i = 1, 2, \dots, k$ a sú nezávislé, tak $\sum_{i=1}^{k} X_i \sim \mathrm{Erl}(k, \lambda)$.

4* Ak $X_i \sim \mathrm{Erl}(n_i, \lambda)$, pre $i = 1, 2, \dots, k$ a sú nezávislé, tak $\sum_{i=1}^{k} X_i \sim \mathrm{Erl}(\sum_{i=1}^{k} n_i, \lambda)$.

5* Ak $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$, pre $i = 1, 2, \dots, k$ a sú nezávislé, tak $\sum_{i=1}^{k} X_i \sim \mathrm{N}(\sum_{i=1}^{k} \mu_i, \sum_{i=1}^{k} \sigma_i^2)$.

3.3.8 Príklad. Predpokladajme, že máme byť obslúžení na dvoch miestach A, B, pričom čas obsluhy v mieste A modeluje veličina X, pričom $X \sim \mathrm{N}(10, 9)$ a čas obsluhy v mieste B modeluje Y, $Y \sim \mathrm{N}(20, 16)$. Určime pravdepodobnosť toho, že súčet časov obsluhy prekročí 35 minút.

Riešenie. Podľa bodu 5 vety 3.3.7 pre $X + Y$ platí: $X + Y \sim \mathrm{N}(30, 25)$. Preto

$$\mathrm{P}(X + Y > 35) = 1 - \mathrm{P}(X + Y < 35) = 1 - F_N\left(\frac{35-30}{5}\right) = 1 - F_N(1) = 1 - 0.84134 = 0.1587.$$

3.3.9 Príklad. V príklade 3.2.6 sme uvažovali o žiarivkách, ktorých životnosť modelovala veličina s rozdelením $\mathrm{Exp}(\lambda)$, kde $\lambda = 0.01$. Predpokladajme, že v prístroji máme jednu žiarivku a naviac sme si pribalili jednu náhradnú. Aká je pravdepodobnosť toho, že počas 300 hodín práce s prístrojom nebudeme mať problém so žiarivkou?

Riešenie. Zrejme ak X, Y modelujú životnosti žiariviek, tak $X, Y \sim \mathrm{Exp}(0.01)$ a sú nezávislé. Preto podľa bodu 3 vety 3.3.7 pre súčet $X + Y$ platí: $X + Y \sim \mathrm{Erl}(2, 0.01)$. Pre hustotu rozdelenia $\mathrm{Erl}(2, \lambda)$, pre $x > 0$, máme $f(x) = \lambda^2 x \exp(-\lambda x)$, kde teraz $\lambda = 0.01$, a preto

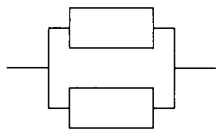
$$\mathrm{P}(X + Y > 300) = 1 - \mathrm{P}(X + Y < 300) = 1 - \int_0^{300} 0.01^2\, x\, e^{-0.01x} dx = 0.20.$$

(Pomôcka: neurčitý integrál funkcie $\lambda^2 x e^{-\lambda x}$ sa rovná $-\lambda x e^{-\lambda x} - e^{-\lambda x}$). Po tomto zistení, keď s prístrojom potrebujeme pracovať najmenej 300 hodín, zrejme pribalíme ešte jednu náhradnú žiarivku (pozri úlohu 3.3.5).

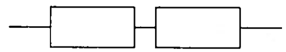
3.3.10 Príklad. Uvažujme o zariadení, ktoré pozostáva z dvoch blokov A, B. Predpokladajme, že životnosti blokov X, Y sú nezávislé veličiny, pre ktoré platí: $X, Y \sim \mathrm{Exp}(\lambda)$. Úloha blokov vzhľadom na zariadenie môže byť

a) buď taká, že zariadenie funguje, ak funguje aspoň jeden z blokov A, B (v kap. 1 sme v takom prípade bloky schematicky znázorňovali ako zapojené vedľa seba, t. j. paralelne),

b) alebo taká, že zariadenie funguje, iba ak obidva bloky fungujú (v takých situáciách sme bloky A, B znázorňovali zapojené za sebou, teda do série).



Obr. 3-2a. Paralelné zapojenie: $Z = max\{X, Y\}$



Obr. 3-2b. Zapojenie do série: $Z = min\{X, Y\}$

Ak veličina Z modeluje životnosť zariadenia, tak v prípade (a) máme $Z = max\{X, Y\}$, kým v prípade (b), máme $Z = min\{X, Y\}$. Nájdime rozdelenie veličiny Z (v oboch prípadoch).

a) V odvodení distribučnej funkcie F_Z veličiny $Z = max\{X, Y\}$ hrá kľúčovú úlohu rovnosť udalostí: $\{max\{X, Y\} < z\} = \{X < z, Y < z\}$ a fakt, že X a Y sú nezávislé. Pre $z < 0$ máme

$$F_Z(z) = \mathrm{P}(max\{X, Y\} < z) = \mathrm{P}(X < z, Y < z) = \mathrm{P}(X < z)\,\mathrm{P}(Y < z) = 0.$$

Ak $z > 0$, tak

$$F_Z(z) = \mathrm{P}(max\{X, Y\} < z) = \mathrm{P}(X < z, Y < z) = \mathrm{P}(X < z)\,\mathrm{P}(Y < z) = (1 - e^{-\lambda z})^2.$$

b) Pri hľadaní F_Z pre $Z = min\{X, Y\}$, využijeme rovnosť $\{min\{X, Y\} > z\} = \{X > z, Y > z\}$.

Pre $z < 0$

$$F_Z(z) = \mathrm{P}(min\{X, Y\} < z) = 1 - \mathrm{P}(min\{X, Y\} > z) = 1 - \mathrm{P}(X > z, Y > z) = 1 - \mathrm{P}(X > z)\,\mathrm{P}(Y > z) = 0,$$

pre $z > 0$

$$F_Z(z) = \mathrm{P}(min\{X, Y\} < z) = 1 - \mathrm{P}(min\{X, Y\} > z) = 1 - \mathrm{P}(X > z, Y > z) =$$
$$= 1 - \mathrm{P}(X > z)\,\mathrm{P}(Y > z) = 1 - e^{-\lambda z} e^{-\lambda z} = 1 - e^{-2\lambda z}.$$

Všimnime si, že rozdelenie $Z = min\{X, Y\}$ je opäť exponenciálne, ale teraz s parametrom 2λ.

3.3.11 Príklad. Nech $X_1, X_2, \ldots, X_n$ sú nezávislé veličiny s rozdelením R(0, 1). Nájdime rozdelenie veličiny $Y = max(X_1, X_2, \ldots, X_n)$.

Riešenie. Pri odvádzaní distribučnej funkcie veličiny Y využijeme, že pre udalosti platí

$$\{Y < y\} = \{max(X_1, X_2, \ldots, X_n) < y\} = \{X_1 < y, X_2 < y, \ldots, X_n < y\}.$$

Zrejme $H(Y) = (0, 1)$, a preto $F_Y(y) = 0$, pre $y \leq 0$, resp. $F_Y(y) = 1$, pre $y \geq 1$. Využívajúc nezávislosť, pre $y \in (0, 1)$ dostávame

$$F_Y(y) = \mathrm{P}(Y < y) = \mathrm{P}(X_1 < y, X_2 < y, \ldots, X_n < y) = \prod_{i=1}^{n} \mathrm{P}(X_i < y) = y^n.$$

Hustotu g rozdelenia veličiny Y získame derivovaním distribučnej funkcie

$$g(y) = \begin{cases} ny^{n-1}, & y \in (0,1) \\ 0, & inde. \end{cases}$$

Tento výsledok využijeme v štatistike pri skúmaní vlastností bodových odhadov (kap. 7).

Úlohy

3.3.1 Určite rozdelenie veličiny Z, keď rozdelenie vektora (X, Y) je dané tabuľkou:

$X \backslash Y$	0	1	2	3	4
0	2/15	1/15	1/15	0	0
1	0	2/15	2/15	1/15	0
2	0	0	2/15	3/15	1/15

a) $Z = X + Y$ c) $Z = XY$ e) $Z = min\{X, Y\}$

b) $Z = X - Y$ d) $Z = |X - Y|$ f) $Z = max\{X, Y\}$

3.3.2 Nech X, Y sú nezávislé. Určite rozdelenie $Z = X + Y$, keď

a) $X \sim \mathrm{Bi}(2, \frac{1}{3})$, $Y \sim \mathrm{Bi}(2, \frac{2}{3})$ b) $X \sim \mathrm{Bi}(2, \frac{1}{3})$, $Y \sim \mathrm{Bi}(3, \frac{1}{3})$

3.3.3 Uvažujme o pokuse, ktorý spočíva v hode tromi mincami, ktorých ruby a líca sme označili číslami 0, 1. Nech X predstavuje číslo, ktoré padne na prvej minci a Y číslo, ktoré je súčtom čísel na druhej a tretej minci. Nakoniec, nech $Z = X + Y$. Určte rozdelenie veličiny Z, keď a) mince sú riadne mince,
b) druhá a tretia minca sú riadne, ale prvá je falošná, pričom $\mathrm{P}(X = 1) = 0.6$.

3.3.4 Modifikujme príklad 3.3.8 tak, že teraz berme do úvahy aj čas potrebný na cestu. Nech Z_1 predstavuje čas potrebný na cestu z domu do miesta A a nech $Z_1 \sim \mathrm{N}(15, 9)$. Ďalej nech Z_2 modeluje čas potrebný na cestu z miesta A do miesta B, pričom $Z_2 \sim \mathrm{N}(20, 16)$ a nakoniec, nech $Z_3 \sim \mathrm{N}(25, 11)$ predstavuje čas na cestu z miesta B domov. Za predpokladu, že všetky uvažované veličiny sú nezávislé, určite pravdepodobnosť toho, že
a) súhrnný čas strávený na ceste prekročí 70 minút,
b) celkový čas potrebný na cestu aj obsluhu neprekročí 100 minút.

3.3.5 V príklade 3.3.9 sme zistili, že $\mathrm{P}(X + Y > 300) = 0.20$, a tak sme sa rozhodli pribaliť ďalšiu, teda druhú rezervnú žiarivku (predpokladajme, že rovnakých parametrov, aké mali tie doteraz uvažované). Určite pravdepodobnosť $\mathrm{P}(X + Y + Z > 300)$, t. j. pravdepodobnosť toho, že teraz tri žiarivky vystačia aspoň na 300 hodín práce.

4 ZÁKLADNÉ ČÍSELNÉ CHARAKTERISTIKY

4.1 Číselné charakteristiky náhodnej veličiny

V tejto kapitole sa dozvieme, ako možno (do istej miery) charakterizovať rozdelenie náhodnej veličiny jednoducho číslom, resp. dvojicou, či trojicou čísel. Ešte vypuklejšie ako doteraz bude evidentné, že s náhodnou veličinou nepracujeme ako so zobrazením, ale to, čo je predmetom nášho záujmu, je jej rozdelenie. Vieme už, že rozdelenie určuje alebo distribučná funkcia, alebo hustota, alebo pravdepodobnostná funkcia. Sú to funkcie reálnej premennej a pre prax sú to pomerne zložité objekty (zadať funkciu je oveľa komplikovanejšie, ako zadať hodnotu, teda zadať číslo). Niekedy však postačuje parciálna informácia o rozdelení. Napr. informácia o tom, kde leží „centrum" hodnôt uvažovanej veličiny, alebo informácia o tom, nakoľko koncentrovane sa realizuje veličina okolo toho centra. To je zmysel číselných charakteristík – poskytnúť parciálnu informáciu o nejakej vlastnosti skúmaného rozdelenia.

Aby sme sa nestratili v záplave nových pojmov, povedzme si, ako možno nazerať na číselné charakteristiky, resp. ako ich môžeme klasifikovať. Jeden uhol pohľadu je, klasifikovať ich podľa ich informačného obsahu. Tak napr. isté charakteristiky hovoria o *polohe* a iné o *rozptýlenosti* (ďalšie o *šikmosti* atď.). Druhý uhol pohľadu je, keď berieme na zreteľ matematické techniky, prostredníctvom ktorých tie charakteristiky počítame. Buď sa pritom oprieme o distribučnú funkciu – v takom prípade hovoríme o *kvantilových* charakteristikách, alebo tie číselné charakteristiky dostaneme operáciou sumovania, alebo integrovania a vtedy hovoríme o *momentových* charakteristikách.

Vzhľadom na naše ciele, obmedzíme sa na základné číselné charakteristiky. Pre prehľadnosť naznačme, o ktorých budeme hovoriť. Sú umiestnené v tabuľke:

	Momentové charakteristiky	Kvantilové charakteristiky
Informácia o polohe	stredná hodnota $\mathrm{E}(X)$	kvantily, resp. medián $med(X)$
Informácia o rozptýlenosti	variancia, resp. smer. odchýlka $var(X)$, resp. $\sigma(X)$	medzikvartilové rozpätie $mkr(X)$

Začneme s charakteristikami polohy. Predtým však urobme dohovor, ktorý bude platiť nielen v tejto, ale aj v ďalších kapitolách. Pod slovami – *rozdelenie veličiny X určuje funkcia f* – rozumieme to, že f je pravdepodobnostná funkcia veličiny X, ak X je diskrétna, resp. že f je hustota X, ak X je spojitá veličina a jej rozdelenie je dané hustotou f.

4.1.1 Definícia. Nech X je náhodná veličina na $(\Omega, \mathcal{S}, \mathrm{P})$ a jej rozdelenie určuje funkcia f.

a) Ak X je diskrétna a $H(X) = \{x_1, x_2, \ldots, x_n\}$, tak stredná hodnota X je číslo

$$\mathrm{E}(X) = \sum_{i=1}^{n} x_i f(x_i)$$

b) Ak X je diskrétna a $H(X)$ je nekonečná, t. j. $H(X) = \{x_1, x_2, \dots, x_n, \dots\}$, tak strednú hodnotu X definujeme ako súčet nekonečného radu

$$E(X) = \sum_{i=1}^{\infty} x_i f(x_i)$$

v prípade, že rad konverguje absolútne (ak rad absolútne nekonverguje, E(X) neexistuje).

c) Ak X je spojitá s hustotou f, tak stredná hodnota je číslo

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$$

v prípade, keď integrál absolútne konverguje (inak hovoríme, že E(X) neexistuje).

4.1.2 Príklad. Nech rozdelenie X je dané tabuľkou. Určme E(X).

x_i	0	1	2	3	4
$f(x_i)$	0.1	0.1	0.2	0.2	0.4

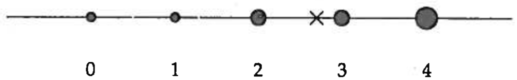
$$E(X) = \sum_{i=1}^{n} x_i f(x_i) = 0{\cdot}0.1 + 1{\cdot}0.1 + 2{\cdot}0.2 + 3{\cdot}0.2 + 4{\cdot}0.4 = 0.1 + 0.4 + 0.6 + 1.6 = 2.7$$

Všimnime si, že stredná hodnota v tomto prípade nie je hodnotou, $E(X) \notin H(X)$. Vezmime na vedomie, že stredná hodnota môže, ale nemusí byť bodom množiny $H(X)$.

4.1.3 V prípade, keď X má konečne veľa hodnôt, na výpočet E(X) môžeme nazerať, ako na výpočet *váženého priemeru* hodnôt x_i

$$E(X) = \sum_{i=1}^{n} x_i f(x_i) = \sum_{i=1}^{n} f(x_i) x_i = \sum_{i=1}^{n} v_i x_i$$

Váhami v_i sú pravdepodobnosti $f(x_i)$. Ďalšiu možnú interpretáciu strednej hodnoty ponúka fyzika. E(X) môžeme interpretovať ako x-ovú súradnicu ťažiska sústavy hmotných bodov, ktorých hmotnosti $f(x_i)$ sú umiestnené na osi x v bodoch so súradnicami x_i. Na obrázku veľkosti bodov predstavujú rôzne hmotnosti $f(x_i)$ z príkladu 4.1.2 a poloha ťažiska je znázornená krížikom. Je dobré mať na mysli túto interpretáciu, lebo niekedy pomôže odhaliť hrubý omyl pri (inak triviálnom) výpočte E(X).



4.1.4 Príklad. (Pravdepodobnostná interpretácia strednej hodnoty) Milan a Cyril hrajú hazardnú hru, v ktorej náhodný mechanizmus realizujú dve (normálne) hracie kocky. Dohodli sa, že Milan vyhráva 10 centov (t. j. Cyril platí 10 centov Milanovi) vtedy, keď na kockách padnú rovnaké čísla. Cyril vyhráva, keď kocky ukážu rôzne čísla. Vtedy Milan platí toľko centov, koľko je rozdiel padnutých bodov na kockách (rozdiel v absolútnej hodnote). Hra je hazardná, pretože náhoda rozhoduje o veľkosti výhry. Aká je pre Cyrila *očakávaná* veľkosť výhry v jednom kole tejto hry?

Riešenie. Hovorme o hre z pozície Cyrila. Výplata (po jednom kole hry) pre Cyrila je náhodná veličina X, ktorej rozdelenie určuje dohoda medzi hráčmi a je dané tabuľkou

x_i	−10	1	2	3	4	5
$f(x_i)$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$

Stredná hodnota X je číslo $E(X) = (-10)\frac{6}{36} + 1\frac{10}{36} + 2\frac{8}{36} + 3\frac{6}{36} + 4\frac{4}{36} + 5\frac{2}{36} = \frac{10}{36} \approx 0.28$

Objasnime, prečo strednú hodnotu môžeme interpretovať ako *očakávanú* hodnotu výhry v jednom kole hry. Predstavme si, že hru si chlapci zahrajú, povedzme 60-krát. Z pohľadu Cyrilových výhier v jednotlivých kolách, to mohlo dopadnúť napríklad takto:

3, 1, 2, 1, -10   4, 1, -10, 3, 4   1, 2, 3, -10, 5   1, 3, 1, 2, 2   1, 4, 4, 3, -10   1, 5, 3, 2, 1
2, -10, 3, 1, 2   3, 1, 2, -10, 4   2, -10, 1, 3, 1   5, 4, 2,-10, 1   2,-10, 1, 3, 1   2, 1,-10, 2, 4

Cyrilova „výhra" -10 (teda vlastne prehra) nastala 10-krát, výhra 1c nastala 17-krát, výhra 2c 13-krát, 3c vyhral Cyril 10 ráz, 4c 7-krát a nakoniec, 5c vyhral 3 razy. Preto celková výhra Cyrila sa rovná

$$(-10){\cdot}10 + 1{\cdot}17 + 2{\cdot}13 + 3{\cdot}10 + 4{\cdot}7 + 5{\cdot}3 = 16, \text{ teda 16 centov.}$$

Keďže celková výhra v 60 kolách sa rovná 16 c, tak (priemerná) výhra na jedno kolo hry sa rovná 16/60, čo je približne 0.27 centov. Všimnime si, nedostali sme 0.28, t. j. E(X), ale to je v poriadku, veď „zapracovala pani Náhoda" a urobili sme len 60 opakovaní pokusu.

Teraz úvahu zopakujme pre n kôl (n opakovaní hry). Celková Cyrilova výhra sa rovná

$$(-10)n_0 + 1n_1 + 2n_2 + 3n_3 + 4n_4 + 5n_5$$

kde n_i sú početnosti jednotlivých výhier. Preto priemerná výhra na jedno kolo (na jednu hru) sa rovná

$$(-10)\frac{n_0}{n} + 1\frac{n_1}{n} + 2\frac{n_2}{n} + 3\frac{n_3}{n} + 4\frac{n_4}{n} + 5\frac{n_5}{n}$$

Ak n je veľké, tak $\frac{n_0}{n} \approx P(X = -10) = f(-10)$ a $\frac{n_i}{n} \approx P(X = i) = f(i)$, pre $i = 1, 2, \dots, 5$. Preto pre veľké n priemerná výhra v jednom kole je blízka hodnote

$$(-10)\, f(-10) + 1f(1) + 2f(2) + 3f(3) + 4f(4) + 5f(5),$$

teda blízka strednej hodnote E(X) veličiny X. To je dôvod, prečo sa niekedy E(X) označuje ako *očakávaná* hodnota veličiny X (symbol E pochádza z anglického slova *expectation*).

4.1.5 Príklad. Uvažujme o náhodnej veličine X, $X \sim Po(\lambda)$. Určime jej strednú hodnotu.

Riešenie. Pripomeňme, že platí

$$f(k) = P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots, n, \dots$$

Hodnoty veličiny X sú prirodzené čísla, a preto absolútna konvergencia radu je to isté ako konvergencia radu a platí

$$E(X) = \sum_{k=0}^{\infty} k\, f(k) = \sum_{k=1}^{\infty} k\, e^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^k}{(k-1)!} = \lambda\, e^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^i}{i!} = \lambda\, e^{-\lambda} e^{\lambda} = \lambda.$$

Ako vidíme, stredná hodnota Poissonovho rozdelenia sa rovná parametru rozdelenia. To je príklad situácie, v ktorej parameter rozdelenia má jednoduchú, presvedčivú interpretáciu. Vo všeobecnosti je E(X) nejakou funkciou parametra. Budeme písať $E(X) = \tau(\theta)$, kde θ je parameter rozdelenia. Pre Poissonove rozdelenie sme zistili, že τ je identita, t. j. $\tau(x) = x$.

4.1.6 Príklad. Nech X má exponenciálne rozdelenie, $X \sim \mathrm{Exp}(\lambda)$. Vypočítajme $\mathrm{E}(X)$ a určime funkciu τ, pre ktorú, v prípade $\mathrm{Exp}(\lambda)$, môžeme písať $\mathrm{E}(X) = \tau(\lambda)$.

Riešenie. Pre hustotu X platí: $f(x) = \lambda\, e^{-\lambda x}$, pre $x > 0$ (inak, $f(x) = 0$). Preto

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_{-\infty}^{0} 0\, dx + \int_{0}^{\infty} x\, \lambda e^{-\lambda x} dx = 0 + \lim_{A\to\infty} \int_0^A \lambda x\, e^{-\lambda x}\, dx =$$

$$= \lambda \lim_{A\to\infty} \left[\frac{-x}{\lambda}\, e^{-\lambda x} - \frac{1}{\lambda^2}\, e^{-\lambda x}\right]_0^A = \lambda\, \frac{1}{\lambda^2} = \frac{1}{\lambda}\,.$$

Integrand $x f(x)$ je funkcia nezáporná, a tak nie je potrebné diskutovať absolútnu konvergenciu integrálu (je totožná s konvergenciou integrálu bez absolútnej hodnoty). Ako vidíme, funkciou $\tau(.)$ je funkcia $\tau(x) = \frac{1}{x}$.

4.1.7 Príklad. Nech $X \sim \mathrm{N}(\mu, \sigma^2)$. Ukážme, že $\mathrm{E}(X) = \mu$.

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (\mu + \sigma t) e^{-\frac{t^2}{2}} \sigma\, dt = \mu\, \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt + 0 = \mu\,.$$

Uplatnili sme substitúciu $t = \frac{x-\mu}{\sigma}$ a potom využili fakt, že funkcia $t e^{-\frac{t^2}{2}}$ je nepárna (a preto konvergentný integrál z tejto funkcie sa na intervale $(-\infty, \infty)$ rovná nule).

4.1.8 Teraz budeme hovoriť o kvantilových charakteristikách polohy. S kvantilmi sme sa stretli v článku 2.3, a teda vieme, že sú to čísla (označujeme ich x_p), ktoré nás informujú o tom, akú pravdepodobnosť má udalosť $\{X < x_p\}$ (platí: $\mathrm{P}(X < x_p) = p$). Rovnako ako v článku 2.3, obmedzíme sa na rozdelenia s hustotou f a naviac také, ktorých distribučná funkcia F je na množine $H(X) = \{x \in \mathrm{R} : f(x) > 0\}$ rastúca (to zaručuje, že F je na $H(X)$ injektívna a nie je nutné pracovať s pseudoinverziou funkcie F). Zopakujme definíciu kvantilu.

4.1.9 Definícia. Nech X je náhodná veličina s hustotou a taká, že jej distribučná funkcia F je na množine $H(X)$ rastúca. Nech $p \in (0, 1)$. Hovoríme, že $x_p \in \mathrm{R}$ je p-kvantil rozdelenia, ktoré má veličina X, ak

$$\mathrm{P}(X < x_p) = p, \ \text{ t.j. } \ F(x_p) = p, \ \text{ t.j. } \ \int_{-\infty}^{x_p} f(x) dx = p.$$

Ak $p = 0.25$, x_p sa nazýva *dolný kvartil*,
ak $p = 0.50$, x_p sa nazýva *medián*,
ak $p = 0.75$, x_p sa nazýva *horný kvartil*.

4.1.10 Príklad. Určme mediány normálneho, rovnomerného a exponenciálneho rozdelenia.

Riešenie. Vieme, že hustota normálneho rozdelenia je funkcia, ktorej graf je symetrický podľa vertikály $x = \mu$, a preto medián sa rovná μ (táto argumentácia sa dá uplatniť pre každé rozdelenie s hustotou, ktorej graf je symetrický podľa vertikály; vtedy stredná hodnota a medián sa rovnajú). Preto máme

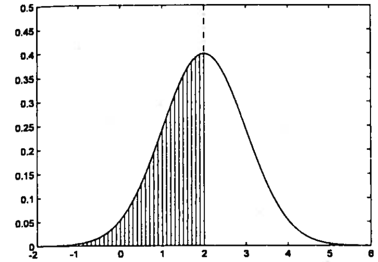
$$\text{ak } X \sim \mathrm{N}(\mu, \sigma^2), \text{ tak } \mathrm{E}(X) = med(X) = \mu\,,$$

$$\text{ak } X \sim \mathrm{R}(a, b), \text{ tak } \mathrm{E}(X) = med(X) = \frac{a+b}{2}.$$

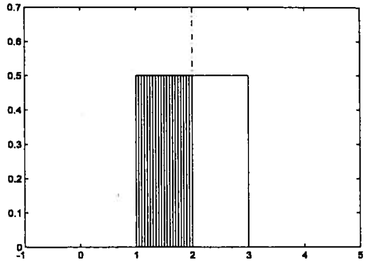
Pre exponenciálne rozdelenie $H(X) = (0, \infty)$ a $F(x) = 1 - e^{-\lambda x}$, pre $x > 0$. Rovnica

$$1 - e^{-\lambda x} = 0.5$$

je ekvivalentná s rovnicou $0.5 = e^{-\lambda x}$ a riešením je medián, $x_{0.5} = med(X) = \frac{\ln 2}{\lambda}$. V príklade 4.1.7 sme zistili, že $\mathrm{E}(X) = \frac{1}{\lambda}$, čo znamená, že teraz platí: $med(X) < \mathrm{E}(X)$.



Obr. 4-1. Ak $X \sim \mathrm{N}(2, 1)$, $\mathrm{E}(X) = med(X) = 2$



Obr. 4-2. Ak $X \sim \mathrm{R}(1, 3)$, $\mathrm{E}(X) = med(X) = 2$

4.1.11 Stredná hodnota a medián predstavujú najdôležitejšie charakteristiky polohy. Predstavujú centrum hodnôt uvažovanej veličiny. Kým $\mathrm{E}(X)$ môžeme interpretovať ako *ťažisko* hodnôt X, medián $med(X)$, je hodnota, ktorá rozdeľuje $H(X)$ na dve časti v tom zmysle, že

$$\mathrm{P}(X < med(X)) = \mathrm{P}(X > med(X)) = \frac{1}{2}$$

Teraz budeme hovoriť o charakteristikách (mierach) rozptýlenosti, ktoré nám poskytujú informáciu o tom, nakoľko tesne sa hodnoty uvažovanej náhodnej veličiny realizujú okolo centra. Základnou momentovou mierou rozptýlenosti je *rozptyl*. Synonymá termínu rozptyl (teda ekvivalentné termíny) sú: *variancia*, resp. *disperzia*. My budeme najčastejšie používať termín variancia.

4.1.12 Definícia. Nech X je náhodná veličina a nech má strednú hodnotu $\mathrm{E}(X)$. Varianciu veličiny X definujeme vzťahom

$$var(X) = \mathrm{E}[(X - \mathrm{E}(X))^2]\,,$$

ak veličina $(X - \mathrm{E}(X))^2$ strednú hodnotu má (ak ju nemá, hovoríme, že $var(X)$ neexistuje). Smerodajnú odchýlku $\sigma(X)$ definujeme vzťahom $\sigma(X) = \sqrt{var(X)}$.

4.1.13 Poznámka. Variancia je stredná hodnota transformovanej veličiny, pričom rozdelenie tej transformácie nepoznáme. Je dobré si uvedomiť, že

$X - \mathrm{E}(X)$	je náhodná veličina, ktorú nazývame *odchýlka* (X od čísla $\mathrm{E}(X)$), rozdelenie odchýlky nie je ťažké určiť, ide o posunutie rozdelenia veličiny X.
$(X - \mathrm{E}(X))^2$	je *kvadratická odchýlka* (odchýlku berieme v druhej mocnine). Táto nová veličina má už zásadne nové rozdelenie a nie je vždy ľahké ho stanoviť. Je zrejmé, že je to nezáporná veličina, ktorá má tým väčšie hodnoty, čím viac sú hodnoty X rozptýlené okolo $\mathrm{E}(X)$.

$\mathrm{E}[(X - \mathrm{E}(X))^2]$ je číslo a to stredná hodnota veličiny $Y = (X - \mathrm{E}(X))^2$, t. j. stredná hodnota kvadratickej odchýlky. Preto sa variancii niekedy hovorí *stredná kvadratická odchýlka*.

Vzhľadom na to, že rozdelenie veličiny $(X - \mathrm{E}(X))^2$ obyčajne nie je jednoduché nájsť, nebudeme $var(X)$ počítať priamo z definície, ale zo vzťahu

$$var(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2,$$

ktorého dôkaz urobíme v článku 4.2. Ako vidíme, potrebujeme zvládnuť výpočet $\mathrm{E}(X^2)$. Ak X je diskrétna veličina, tak to nie je žiaden problém, ale prípad spojitej X s hustotou, si vyžaduje pozornosť (pozri odsek 4.1.15).

Na záver tejto poznámky objasnime, prečo má vôbec zmysel hovoriť o smerodajnej odchýlke $\sigma(X)$, keď je definovaná tak jednoduchým vzťahom: $\sigma(X) = \sqrt{var(X)}$. Sú na to aspoň dva dôvody. Ak X má fyzikálny rozmer, tak rozmer $var(X)$ je už iný (väčšinou ťažko interpretovateľný), ale rozmer $\sqrt{var(X)}$ sa zhoduje s rozmerom X, a preto je možné porovnávať $\mathrm{E}(X)$ so $\sigma(X)$. Druhý dôvod – ak X má rozdelenie $\mathrm{N}(\mu, \sigma^2)$, tak $var(X) = \sigma^2$, čo znamená, že smerodajná odchýlka σ má význam parametra. Naviac, body $\mu - \sigma$, $\mu + \sigma$ sú súradnice inflexných bodov Gaussovej krivky (teda významných bodov grafu hustoty).

4.1.14 Príklad. Nech rozdelenie X je dané tabuľkou. Určme $var(X)$.

x_i	0	1	2	3	4
$f(x_i)$	0.1	0.1	0.2	0.2	0.4

Budeme postupovať podľa výpočtového vzťahu z 4.1.13. V príklade 4.1.2 sme vypočítali $\mathrm{E}(X)$, a teda vieme, že $\mathrm{E}(X) = 2.7$. Z tabuľky pravdepodobnostnej funkcie $Y = X^2$

y_j	0	1	4	9	16
$g(y_j)$	0.1	0.1	0.2	0.2	0.4

máme $\mathrm{E}(X^2) = 0.1 + 0.8 + 1.8 + 6.4 = 9.1$ a $var(X) = \mathrm{E}(X^2) - (\mathrm{E}X)^2 = 9.1 - (2.7)^2 = 1.81$.

4.1.15 Príklad. Nech $X \sim \mathrm{R}(-2, 2)$. Určime $var(X)$.

Riešenie. Opäť postupujeme podľa výpočtového vzťahu $var(X) = \mathrm{E}(X^2) - (\mathrm{E}X)^2$ a keďže podľa 4.1.10 máme $\mathrm{E}(X) = 0$, ide o výpočet $\mathrm{E}(X^2)$. Potrebujeme teda nájsť hustotu g veličiny $Y = X^2$. Tú sme našli v príklade 2.5.9 a dostali sme

$$g(y) = \begin{cases} \dfrac{1}{4\sqrt{y}}, & 0 < y < 4 \\ 0, & inak. \end{cases}$$

Preto

$$var(X) = \mathrm{E}(X^2) = \mathrm{E}(Y) = \int_{-\infty}^{\infty} y g(y) dy = \int_0^4 y \frac{1}{4\sqrt{y}}\, dy = \frac{1}{4} \int_0^4 \sqrt{y}\, dy = \frac{1}{4} \left[\frac{2}{3} y^{\frac{3}{2}}\right]_0^4 = \frac{4}{3}.$$

Je fakt, že nájsť hustotu transformácie $Y = X^2$ nie je vždy jednoduché a hlavne, niekedy nás nezaujíma rozdelenie X^2, len výpočet $var(X)$. Našťastie matematická analýza nám ponúka tvrdenie, pomocou ktorého na výpočet $\mathrm{E}(X^2)$ nebudeme potrebovať hustotu $Y = X^2$.

4.1.16 Veta. Nech X je veličina, ktorej rozdelenie určuje funkcia f. Nech $h\colon \mathrm{R} \to \mathrm{R}$ je taká, že veličina $h(X)$ má strednú hodnotu. Potom

$$\mathrm{E}(h(X)) = \textstyle\sum_i h(x_i) f(x_i),$$

ak X je diskrétna a $H(X) = \{x_1, x_2, x_3, \dots\}$, resp.

$$\mathrm{E}(h(X)) = \int_{-\infty}^{\infty} h(x) f(x)\, dx,$$

ak X je spojitá s hustotou f.

4.1.17 Príklad. Nech $X \sim \mathrm{R}(-2, 2)$. Určme $var(X)$, tentoraz použitím vety 4.1.16.

Riešenie. Podľa výpočtového vzťahu $var(X) = \mathrm{E}(X^2) - (\mathrm{E}X)^2 = \mathrm{E}(X^2)$, pretože $\mathrm{E}(X) = 0$. Funkciou h je $h(x) = x^2$ a použitie 4.1.16 dáva

$$var(X) = \mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_{-2}^{2} \frac{1}{4} x^2 dx = \frac{1}{4} \int_{-2}^{2} x^2 dx = \frac{1}{12} [x^3]_{-2}^{2} = \frac{16}{12} = \frac{4}{3}.$$

Ako bolo povedané v úvode článku, variancia (rozptyl) je momentovou charakteristikou rozptýlenosti. Nasledujúca definícia uvádza ďalšiu možnú mieru rozptýlenosti, ktorá je kvantilovou charakteristikou (lebo sa opiera o kvantily).

4.1.18 Definícia. Nech X je náhodná veličina s hustotou a taká, že jej distribučná funkcia F je na množine $H(X)$ rastúca. Medzikvartilové rozpätie $mkr(X)$ definujeme vzťahom

$$mkr(X) = x_{0.75} - x_{0.25}$$

t. j. $mkr(X)$ je vzdialenosť medzi dolným a horným kvartilom rozdelenia veličiny X. Samozrejme, čím je rozptýlenosť veličiny X väčšia, tým väčšie je $mkr(X)$.

4.1.19 Príklad. Nech $X \sim \mathrm{Exp}(\lambda)$. Určme $mkr(X)$.

Riešenie. Je potrebné stanoviť hodnoty dolného a horného kvartilu, teda $x_{0.25}$, resp. $x_{0.75}$. Vieme, že sú to riešenia rovníc

$$1 - e^{-\lambda x} = 0.25, \quad \text{resp.} \quad 1 - e^{-\lambda x} = 0.75$$

Riešením prvej rovnice je $x = \frac{\ln\frac{4}{3}}{\lambda}$, t. j. $x_{0.25} = \frac{\ln\frac{4}{3}}{\lambda}$, riešením druhej, $x = \frac{\ln 4}{\lambda}$, t. j. $x_{0.75} = \frac{\ln 4}{\lambda}$. Takto pre $mkr(X)$ máme

$$mkr(X) = \frac{\ln 4}{\lambda} - \frac{\ln\frac{4}{3}}{\lambda} = \frac{\ln 3}{\lambda}$$

4.1.20 Veta (Čebyševova nerovnosť). Nech X je ľubovoľná taká, že existujú $\mathrm{E}(X)$ a $var(X)$. Potom pre každé kladné číslo ε platí nerovnosť

$$\mathrm{P}(\,|X - \mathrm{E}(X)| \geq \varepsilon\,) \leq \frac{var(X)}{\varepsilon^2}$$

Nerovnosť hovorí, že pravdepodobnosť udalosti, ktorá spočíva v tom, že X sa realizuje mimo ε okolia $\mathrm{E}(X)$ je zhora ohraničená zlomkom na pravej strane nerovnosti. Čím je $var(X)$ menšia, tým je pravdepodobnosť udalosti $\{|X - \mathrm{E}(X)| \geq \varepsilon\}$ menšia. Teda naozaj, $var(X)$ je miera rozptýlenosti. Všimnime si, že keď za ε berieme malé číslo (napr. 0.1), tak ε^2 je veľmi malé (napr. 0.01), a pretože je v menovateli, pravá strana môže mať aj hodnotu väčšiu ako 1. V takom prípade, samozrejme, nerovnosť nič nehovorí. Je zaujímavé, že nerovnosť platí za veľmi širokých predpokladov (tak pre diskrétne, ako aj spojité veličiny).

Úlohy

4.1.1 Vezmime na vedomie (resp. pre niektoré rozdelenia overte), že platí:

$X \sim \mathrm{A}(p)$	$\mathrm{E}(X) = p$	$var(X) = p(1-p)$
$X \sim \mathrm{R}\{x_1, x_2, \dots, x_n\}$	$\mathrm{E}(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$	$var(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\mathrm{E}X)^2$
$X \sim \mathrm{Bi}(n, p)$	$\mathrm{E}(X) = np$	$var(X) = np(1-p)$
$X \sim \mathrm{H}(N, M, n)$	$\mathrm{E}(X) = n\frac{M}{N}$	$var(X) = n\frac{M}{N}\left(1 - \frac{M}{N}\right)\left(1 - \frac{n-1}{N-1}\right)$
$X \sim \mathrm{Po}(\lambda)$	$\mathrm{E}(X) = \lambda$	$var(X) = \lambda$
$X \sim \mathrm{G}(p)$	$\mathrm{E}(X) = \frac{1-p}{p}$	$var(X) = \frac{1-p}{p^2}$
$X \sim \mathrm{R}(a, b)$	$\mathrm{E}(X) = \frac{a+b}{2}$	$var(X) = \frac{(b-a)^2}{12}$
$X \sim \mathrm{Exp}(\lambda)$	$\mathrm{E}(X) = \frac{1}{\lambda}$	$var(X) = \frac{1}{\lambda^2}$
$X \sim \mathrm{Erl}(k, \lambda)$	$\mathrm{E}(X) = \frac{k}{\lambda}$	$var(X) = \frac{k}{\lambda^2}$
$X \sim \mathrm{N}(\mu, \sigma^2)$	$\mathrm{E}(X) = \mu$	$var(X) = \sigma^2$

4.1.2 Pre niektoré úlohy z úloh 2.1.1 až 2.1.20 určite $\mathrm{E}(X)$ a $var(X)$.

4.1.3 Rozhodnite, ktorá z veličín X, Y má menšiu varianciu, keď rozdelenie X je rovnomerné $\mathrm{R}\{-2, -1, 0, 1, 2\}$ a Y má rozdelenie dané tabuľkou

y_i	-3	-2	-1	0	1	2	3
$\mathrm{P}(Y = y_i)$	0.05	0.10	0.15	0.40	0.15	0.10	0.05

4.1.4 Pre veličiny Y, Z, U, V z úlohy 2.5.1 určte ich stredné hodnoty a variancie. Preverte výsledky výpočtov tak, že budete postupovať dvomi spôsobmi – jednak podľa definície $\mathrm{E}(X)$, resp. $\mathrm{E}(X^2)$ (využívajúc výsledky úlohy 2.5.1) a jednak použitím vety 4.1.16.

4.1.5 Ukážte, že pre náhodnú veličinu z príkladu 2.1.6 stredná hodnota $\mathrm{E}(X)$ neexistuje.

4.1.6 Pre niektoré úlohy z úloh 2.2.1 až 2.2.12 určite $\mathrm{E}(X)$ a $var(X)$, $med(X)$ a $mkr(X)$.

4.1.7 Nech $X \sim \mathrm{R}(-a, a)$, $Y \sim \Delta(-b, 0, b)$, $Z \sim \mathrm{N}(c, d^2)$. Nájdite hodnoty parametrov a, b, c, d tak, aby $\mathrm{E}(X) = \mathrm{E}(Y) = \mathrm{E}(Z) = 0$, a pritom aby $var(X) = var(Y) = var(Z) = 1$. Úloha ilustruje základnú skutočnosť, že stredná hodnota a variancia zďaleka neurčujú (vo všeobecnosti) rozdelenie náhodnej veličiny.

4.1.8 Je dané rozdelenie X a nech $Y = h(X)$. Nájdite $\mathrm{E}(Y)$ použitím vety 4.1.16. Potom overte správnosť výsledku výpočtom $\mathrm{E}(Y)$ z definície.

a) Nech $X \sim \mathrm{R}(0, 2)$ a $Y = 1 + 2X$ (rozdelenie Y sme našli v príklade 2.5.5).

b) Nech $X \sim \Delta(0, 2, 2)$ a $Y = 3 + \frac{X}{2}$ (rozdelenie Y sme našli v príklade 2.5.6).

c) Nech $X \sim \Delta(0, 1, 1)$ a $Y = X^2$ (o rozdelení Y hovorí úloha 2.5.14).

4.2 Číselné charakteristiky náhodného vektora

Ak (X, Y) je náhodný vektor, tak jeho zložky vieme charakterizovať tak, ako sme to ukázali v predchádzajúcom článku. Z kapitoly 3 však vieme, že v prípade vektora nás zaujímajú nielen zložky X, Y samé o sebe (resp. ich rozdelenia), ale aj stochastická väzba, *asociácia* medzi zložkami. V tomto článku budeme hovoriť o mierach, t. j. o číselných charakteristikách, ktoré opisujú asociáciu medzi zložkami vektora, teda charakterizujú väzbu medzi veličinami X, Y. Oprieme sa o vetu 4.2.1, ktorá má kľúčovú úlohu jednak pri výpočtoch týchto charakteristík a tiež pri dokazovaní ich vlastností.

Nech (X, Y) je náhodný vektor a $h\colon \mathrm{R} \times \mathrm{R} \to \mathrm{R}$ je funkcia predstavujúca transformáciu zložiek vektora. V článku 3.3 sme sa zapodievali hľadaním rozdelenia veličiny $Z = h(X, Y)$. Tak napríklad, hľadali sme rozdelenie veličiny $Z = X + Y$, alebo $Z = X\,Y$. Uvedomili sme si, že (hlavne v spojitom prípade) nájsť rozdelenie veličiny Z vôbec nemusí byť jednoduché. Niekedy nám však stačí parciálna informácia, napríklad, potrebujeme poznať $\mathrm{E}(Z)$, resp. $var(Z)$, a preto vítame, že máme k dispozícii vetu 4.2.1, ktorá umožňuje výpočet $\mathrm{E}(h(X, Y))$ bez znalosti rozdelenia veličiny $Z = h(X, Y)$. Ide o dvojrozmernú analógiu vety 4.1.16.

4.2.1 Veta. Nech (X, Y) je náhodný vektor s rozdelením, ktoré určuje funkcia $f(.,.)$. Nech $h\colon \mathrm{R} \times \mathrm{R} \to \mathrm{R}$ je taká funkcia, že veličina $h(X, Y)$ má strednú hodnotu. Potom platí

$$\mathrm{E}(h(X, Y)) = \sum_i \sum_j h(x_i, y_j) f(x_i, y_j),$$

ak (X, Y) je diskrétny vektor s pravdepodobnostnou funkciou f. Ak (X, Y) je vektor s hustotou f, tak platí

$$\mathrm{E}(h(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) f(x, y)\, dx\, dy.$$

4.2.2 Príklad. Náhodný vektor má rozdelenie dané tabuľkou pravdepodobnostnej funkcie

(X, Y)	0	1	2	3	4
0	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{2}{24}$	0	0
1	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{2}{24}$	$\frac{2}{24}$	0
2	0	$\frac{1}{24}$	$\frac{3}{24}$	$\frac{2}{24}$	$\frac{1}{24}$
3	0	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{1}{24}$

Nájdime $\mathrm{E}(X\,Y)$ najprv bez použitia vety 4.2.1 a potom s jej využitím. Postupovať bez použitia vety 4.2.1 znamená nájsť pravdepodobnostnú funkciu g veličiny $Z = X\,Y$ a potom počítať $\mathrm{E}(Z)$ z definície strednej hodnoty. Funkcia g je daná tabuľkou:

z_i	0	1	2	3	4	6	8	9	12
$g(z_i)$	$\frac{6}{24}$	$\frac{2}{24}$	$\frac{3}{24}$	$\frac{3}{24}$	$\frac{3}{24}$	$\frac{3}{24}$	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{1}{24}$

Trochu práce dalo nájsť rozdelenie veličiny Z, avšak výpočet $E(Z)$ je už celkom štandardný

$$E(X\,Y) = E(Z) = 0 + \frac{2}{24} + \frac{6}{24} + \frac{9}{24} + \frac{12}{24} + \frac{18}{24} + \frac{8}{24} + \frac{18}{24} + \frac{12}{24} = \frac{85}{24} = 3.541\bar{6} \approx 3.5417.$$

A teraz vypočítajme $E(X\,Y)$ použitím vety 4.2.1. Prechádzame bunky tabuľky, napríklad po riadkoch (zľava doprava a zhora dole), uvažujúc hodnoty $h(x_i, y_j) = x_i y_j$ a obsah buniek tabuľky, t.j. pravdepodobnosti $f(x_i, y_j)$. Nulové členy nepíšeme, a preto začíname až druhou bunkou v druhom riadku tabuľky

$$E(X\,Y) = 1\frac{2}{24} + 2\frac{2}{24} + 3\frac{2}{24} + 2\frac{1}{24} + 4\frac{3}{24} + 6\frac{2}{24} + 8\frac{1}{24} + 3\frac{1}{24} + 6\frac{1}{24} + 9\frac{2}{24} + 12\frac{1}{24} = \frac{85}{24}.$$

Pri výpočtoch má veta 4.2.1 význam hlavne pre spojitý vektor s hustotou (to je však prípad, ktorému sa nemôžeme venovať). Ako sme už povedali, okrem výpočtov je veta 4.2.1 dôležitá preto, lebo má kľúčovú úlohu pri dokazovaní vlastností operátora E strednej hodnoty. Na symbol E môžeme totiž nazerať ako na zobrazenie (funkcionál), ktorého definičný obor je množina všetkých náhodných veličín na danom pravdepodobnostnom priestore (pre ktoré stredná hodnota existuje). Obor hodnôt funkcionálu E je množina reálnych čísel. Vzhľadom na vlastnosti uvedené vo vete 4.2.3 hovoríme, že E predstavuje *pozitívny lineárny funkcionál.*

4.2.3 Veta. (Vlastnosti strednej hodnoty) Nech X, Y sú náhodné veličiny na tom istom pravdepodobnostnom priestore a majú stredné hodnoty, teda existujú $E(X)$, $E(Y)$. Potom platí:

1. Ak $a, b \in R$, tak $E(aX + bY) = aE(X) + bE(Y)$.
2. Ak $X \geq 0$, tak $E(X) \geq 0$.
3. Ak $a, b \in R$, tak $E(a + bX) = a + bE(X)$.
4. Ak $X \leq Y$, tak $E(X) \leq E(Y)$.
5. Ak X a Y sú nezávislé, tak $E(X\,Y) = E(X)\,E(Y)$.

Základnými sú vlastnosti 1, 2, 5, bod 3 je dôsledok bodu 1 a bod 4 je dôsledkom 1 a 2. Na reálne číslo a môžeme nazerať ako na degenerovanú (náhodnú) veličinu – ako na diskrétnu veličinu, ktorá má jedinú hodnotu (ňou je číslo a), ktorú nadobúda s jednotkovou pravdepodobnosťou. Zrejme potom $E(a) = a$ a je jasné, že bod 3 je dôsledkom bodu 1.

Ukážme, ako využijeme vetu 4.2.1 v dôkaze, napríklad, prvého tvrdenia vety. Máme dokázať, že operátor strednej hodnoty je lineárny operátor (to je obsah bodu 1). Na veličinu $aX + bY$ nazeráme ako na $h(X,Y)$, keď $h(x,y) = ax + by$. Preto v diskrétnom prípade máme

$$E(aX + bY) = \sum_i \sum_j h(x_i, y_j) f(x_i, y_j) = \sum_i \sum_j (ax_i + by_j) f(x_i, y_j) = \sum_i \sum_j ax_i f(x_i, y_j) +$$
$$+ \sum_j \sum_i by_j f(x_i, y_j) = \sum_i ax_i f_1(x_i) + \sum_j by_j f_2(y_j) = a\sum_i x_i f_1(x_i) + b\sum_j y_j f_2(y_j).$$

Analogicky sa dokáže bod 5. Teraz však transformáciou bude $h(x,y) = xy$ a vďaka nezávislosti X a Y, využijeme rovnosť

$$f(x_i, y_j) = f_1(x_i) f_2(y_j)$$

pre všetky $x_i \in H(X)$, $y_j \in H(Y)$. Veta 4.2.3 je jedna z tých, na ktoré sa budeme v ďalších kapitolách veľmi často odvolávať.

4.2.4 Príklad. Nech $X \sim N(\mu, \sigma^2)$ a $Y = a + bX$. Určíme $E(Y)$.

Riešenie. Doteraz sme $E(Y)$ určovali podľa vety 2.5.7 (tá veta nám hovorí nielen o $E(Y)$, ale aj o tom, že rozdelenie Y je normálne). Pokiaľ však ide len o určenie $E(Y)$, jednoduchšie je odvolať sa na vetu 4.2.3 a počítať $E(Y)$ takto:

$$E(Y) = E(a + bX) = E(a) + bE(X) = a + b\mu.$$

Pretože variancia X je stredná hodnota kvadratickej odchýlky, $var(X) = E[(X - EX)^2]$, je jasné, že vlastnosti operátora E sa prejavia na vlastnostiach variancie. Práve sme použili označenie $E[(X - EX)^2]$ namiesto presného $E[(X - E(X))^2]$ – toto zjednodušenie, keď namiesto $E(X)$ píšeme EX, budeme používať, v snahe zľahčiť čitateľnosť niektorých výrazov.

4.2.5 Veta (Vlastnosti variancie) Nech X, Y sú náhodné veličiny na tom istom pravdepodobnostnom priestore a existujú $var(X)$, $var(Y)$. Potom platí:

1. $var(X) \geq 0$, pre každú náhodnú veličinu X.
2. $var(aX) = a^2\, var(X)$, pre každé $a \in R$.
3. $var(X + c) = var(X)$, pre každé $c \in R$.
4. $var(X + Y) = var(X) + var(Y) + 2\,E[(X - EX)(Y - EY)]$.
5. Ak X, Y sú nezávislé, tak $var(X \pm Y) = var(X) + var(Y)$.

Dôkaz. Pretože $(X - EX)^2 \geq 0$, podľa bodu 2 vety 4.2.3. zrejme $E[(X - EX)^2] \geq 0$. Body 2, resp. 3 sa dokazujú ľahko, urobme dôkaz 4 a 5. Zrejme

$$var(X + Y) = E[(X + Y - E(X + Y))^2] = E[\,((X - EX) + (Y - EY))^2\,] =$$
$$= E[(X - EX)^2 + 2(X - EX)(Y - EY) + (Y - EY)^2\,] = var(X) + var(Y) + 2E[(X - EX)(Y - EY)].$$

Dôkaz bodu 5: Ukážme, že pre nezávislé X, Y výraz $E[(X - EX)(Y - EY)] = 0$. Vskutku,

$$E[(X - EX)(Y - EY)] = E[\,X\,Y - E(X)Y - E(Y)X + E(X)E(Y)\,] = E(X\,Y) - E(X)E(Y) -$$
$$- E(Y)E(X) + E(X)E(Y) = E(X\,Y) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0,$$

pretože podľa bodu 5 vety 4.2.3 pre nezávislé X, Y máme $E(X\,Y) = E(X)E(Y)$.

4.2.6 Príklad. Nech $X \sim N(\mu, \sigma^2)$ a $Y = a + bX$. Určíme $var(Y)$.

Riešenie. Ide o analogickú situáciu ako v príklade 4.2.4. Nie je nutné odvolávať sa na vetu 2.5.7, ale určiť $var(Y)$ môžeme takto

$$var(Y) = var(a + bX) = var(bX) = b^2 var(X) = b^2\sigma^2.$$

4.2.7 Veta (Normovanie náhodnej veličiny). Ak veličina X má $E(X)$ a $var(X) > 0$, tak platí:

$$\text{Ak } X^o = \frac{X - E(X)}{\sigma(X)}, \text{ potom } E(X^o) = 0,\ var(X^o) = 1.$$

Dôkaz sa urobí postupmi z odsekov 4.2.4 a 4.2.6, veď predpoklad normality v nich nehral žiadnu úlohu. Veličina X^o je lineárna transformácia X, $X^o = a + bX$, kde $a = -\frac{E(X)}{\sigma(X)}$, $b = \frac{1}{\sigma(X)}$. Proces normovania sme fakticky realizovali už v článku 2.5, avšak tam sme pracovali s normálnym rozdelením, kým v tejto vete žiadne konkrétne rozdelenie nepredpokladáme.

4.2.8 Príklad. Nech $X \sim \mathrm{Erl}(k, \lambda)$. Určíme a, b tak, aby pre $Y = a + bX$ platilo:

$$\mathrm{E}(Y) = 0, \ var(Y) = 1.$$

Nájdime konkrétne hodnoty a, b, ak $k = 4$, $\lambda = 0.1$.

Riešenie. Z toho, čo sme uviedli v 4.2.7 je zrejmé, že za Y stačí zobrať X^o, to znamená, že $a = -\frac{E(X)}{\sqrt{var(X)}} = -\frac{\frac{k}{\lambda}}{\sqrt{\frac{k}{\lambda^2}}} = -\sqrt{k}$, $b = \frac{1}{\sqrt{\frac{k}{\lambda^2}}} = \frac{\lambda}{\sqrt{k}}$. Ak $X \sim \mathrm{Erl}(4, 0.1)$, tak $Y = -2 + 0.05X$ je taká, že $\mathrm{E}(Y) = 0$ a $var(Y) = 1$ (overte, uplatňujúc vety 4.2.3, 4.2.5). Teda $a = -2$, $b = 0.05$.

Vlastnosť strednej hodnoty z bodu 1 vety 4.2.3 a vlastnosti variancie z bodov 4 a 5 vety 4.2.5 majú prirodzené zovšeobecnenie pre konečne veľa náhodných veličín. V nasledujúcej vete sa venujeme prípadu, v ktorom uvažované veličiny sú nezávislé, pretože práve taká situácia bude v ďalších kapitolách najčastejšia.

4.2.9 Veta. Ak X_i, $i = 1, 2, \ldots, n$ sú nezávislé veličiny a majú variancie, tak platí

$$\mathrm{E}(X_1 + X_2 + \ldots + X_n) = \mathrm{E}(X_1) + \mathrm{E}(X_2) + \ldots + \mathrm{E}(X_n),$$

$$var(X_1 + X_2 + \ldots + X_n) = var(X_1) + var(X_2) + \ldots + var(X_n).$$

4.2.10 Príklad. Hracou kockou hoďme 12-krát a nech X predstavuje počet 6-tiek, ktoré padli v tých 12 hodoch. Stanovme $\mathrm{E}(X)$ a $var(X)$.

Riešenie. Už vieme, že $X \sim \mathrm{Bi}(12, 1/6)$. Keď využijeme výsledky úlohy 4.1.1, dostávame $\mathrm{E}(X) = 12(1/6) = 2$ a $var(X) = 12(1/6)(5/6) = 10/6$. Môžeme však postupovať aj tak, že s i-tým hodom spojíme predstavu veličiny X_i, kde X_i je indikátor čísla 6 v i-tom hode. Zrejme X_i sú nezávislé, $X_i \sim \mathrm{A}(1/6)$ a $X = X_1 + X_2 + \ldots + X_{12}$. Aplikovaním vety 4.2.9 získame rovnaké výsledky, pretože pre všetky $i = 1, 2, \ldots, 12$ zrejme platí: $\mathrm{E}(X_i) = 1/6$, $var(X_i) = 5/36$.

Teraz sa venujme výrazu $\mathrm{E}[(X - \mathrm{E}X)(Y - \mathrm{E}Y)]$, ktorý vystupuje v bode 4 vety 4.2.5. Ide o strednú hodnotu veličiny $(X - \mathrm{E}X)(Y - \mathrm{E}Y)$, ktorá – ako ukážeme – do istej miery opisuje asociáciu medzi X a Y.

4.2.11 Definícia. Nech X, Y sú náhodné veličiny na tom istom pravdepodobnostnom priestore a existujú $var(X)$, $var(Y)$. Číslo

$$cov(X, Y) = \mathrm{E}[(X - \mathrm{E}X)(Y - \mathrm{E}Y)]$$

nazývame *kovariancia* veličín X, Y. Ak $var(X) > 0$, $var(Y) > 0$, tak *korelačný koeficient* definujeme vzťahom

$$\rho(X, Y) = \frac{\mathrm{E}[(X - \mathrm{E}X)(Y - \mathrm{E}Y)]}{\sqrt{var(X)}\sqrt{var(Y)}}.$$

Ak $\rho(X, Y) = 0$, tak veličiny X, Y nazývame *nekorelované*.

4.2.12 Poznámka. Výpočet kovariancie z definície nie je praktický. Jednoduchšie je postupovať podľa vzťahu

$$cov(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y),$$

ktorého platnosť sme ukázali v rámci dôkazu bodu 5 vety 4.2.5. Ilustrujme jeho použitie pre vektor (X, Y) z príkladu 4.2.2.

$$cov(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = \frac{85}{24} - 1.5 \cdot 2 = \frac{85-72}{24} = \frac{13}{24} = 0.541\overline{6} \approx 0.5417.$$

4.2.13 Poznámka. Už vieme, ako je kovariancia definovaná a vieme ju vypočítať. Avšak zatiaľ sme neukázali, že dáva (do istej miery) informáciu o asociácii medzi X a Y. Hovoríme, že veličiny X, Y sú *pozitívne asociované*, keď veľké hodnoty X korešpondujú s veľkými hodnotami Y a malé hodnoty X s malými hodnotami Y (ako malé označujeme tie, čo ležia naľavo od strednej hodnoty, ako veľké tie, čo ležia napravo od nej). Teda pozitívna asociovanosť znamená, že veličina $(X - \mathrm{E}X)(Y - \mathrm{E}Y)$ je (s veľkou pravdepodobnosťou) nezáporná, a preto jej stredná hodnota, teda $cov(X, Y) = \mathrm{E}[(X - \mathrm{E}X)(Y - \mathrm{E}Y)]$, je číslo nezáporné (odvolávame sa na bod 2 vety 4.2.3). Pri negatívnej asociovanosti je to naopak.

4.2.14 Príklad. Uvažujme o vektore (X, Y) z príkladu 4.2.2.

$X \backslash Y$	0	1	2	3	4
0	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{2}{24}$	0	0
1	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{2}{24}$	$\frac{2}{24}$	0
2	0	$\frac{1}{24}$	$\frac{3}{24}$	$\frac{2}{24}$	$\frac{1}{24}$
3	0	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{1}{24}$

Za malé hodnoty X považujeme 0 a 1, tie korešpondujú s hodnotami 0, 1, 2 veličiny Y, ktoré môžeme chápať ako malé (až na hodnotu 2, pretože tá sa rovná strednej hodnote Y). Za veľké hodnoty X považujeme 2, 3, ktoré korešpondujú s veľkými hodnotami Y – okrem hodnoty 1, ktorá je zrejme malá. Avšak, pravdepodobnosť udalosti $(X, Y) \in \{(2, 1), (3, 1)\}$ je malá v porovnaní s tou, ktorá odpovedá výsledkom, ktoré podporujú tvrdenie:

Veľké hodnoty X korešpondujú s veľkými hodnotami Y a malé hodnoty X s malými hodnotami Y. O takejto skutočnosti nás informuje kovariancia $cov(X, Y)$ svojou kladnou hodnotou $0.541\overline{6}$ (pozri poznámku 4.2.10).

Uvážme, že keby bola kovariancia záporná, išlo by o negatívnu asociáciu a v takom prípade by bunky tabuľky s kladnými pravdepodobnosťami ležali (voľne povedané) nie v smere hlavnej diagonály (ako je to v našej tabuľke), ale v smere vedľajšej diagonály.

4.2.15 Príklad. Nech (X, Y) je vektor z predchádzajúceho príkladu a uvažujme o vektore (U, V), pre ktorý platí $U = 4X$, $V = 3Y$. Ukážme, že $cov(U, V) = 12\,cov(X, Y)$.

Riešenie. Jedna možnosť ako ukázať platnosť rovnosti spočíva v tom, že nájdeme tabuľku pravdepodobnostnej funkcie vektora (U, V) a z nej vypočítame $cov(U, V)$. Nie je to šikovné riešenie, ale je zaiste poučné hľadať tabuľku pravdepodobnostnej funkcie vektora (U, V).

Elegantnejšia cesta je nájsť $cov(U, V)$, využívajúc vlastnosti operátora E (veta 4.2.3)

$$cov(U, V) = \mathrm{E}[(U - \mathrm{E}U)(V - \mathrm{E}V)] = \mathrm{E}[\,(4X - \mathrm{E}(4X))\,(3Y - \mathrm{E}(3Y))\,] = \mathrm{E}[(4X - 4\,\mathrm{E}X)(3Y - 3\,\mathrm{E}Y)] =$$

$$= \mathrm{E}[4(X - \mathrm{E}X)\,3(Y - \mathrm{E}Y)\,] = 12\,\mathrm{E}[(X - \mathrm{E}X)(Y - \mathrm{E}Y)] = 12cov(X, Y).$$

4.2.16 Poznámka. Z príkladu 4.2.15 je vidieť, že platí: $cov(aX,\ bY) = ab\ cov(X,Y)$, čo znamená, že zmenou mierky (ktorú uplatňujeme na hodnoty veličín X,Y) sa mení hodnota kovariancie. To je dôvod, prečo $cov(X,Y)$ nepovažujeme za vhodnú mieru závislosti X,Y. K. Pearson navrhol za mieru závislosti medzi X a Y, vziať kovarianciu *normovaných* X,Y, a tak vlastne vznikol *korelačný koeficient* $\rho(X,Y)$. Vskutku, ukážeme, že platí

$$\rho(X,Y) = cov(X^o, Y^o).$$

Začnime úpravou pravej strany: $cov(X^o,Y^o) = \mathrm{E}(X^oY^o) - \mathrm{E}(X^o)\mathrm{E}(Y^o) = \mathrm{E}(X^oY^o) =$

$$= \mathrm{E}\left[\frac{X-EX}{\sqrt{var(X)}}\,\frac{Y-EY}{\sqrt{var(Y)}}\right] = \frac{1}{\sqrt{var(X)}\sqrt{var(Y)}}\ \mathrm{E}[(X-\mathrm{E}X)(Y-\mathrm{E}Y)] = \rho(X,Y).$$

Presnejšie povedané, vzhľadom na bod 2 vety 4.2.17, korelačný koeficient považujeme za mieru *lineárnej závislosti* medzi X a Y.

4.2.17 Veta. (Vlastnosti korelačného koeficienta) Nech $X,\ Y$ sú náhodné veličiny (na tom istom pravdepodobnostnom priestore) a majú kladné variancie. Potom

1. $\rho(X,Y) = \rho(Y,X)$ a pre $\rho(X,Y)$ vždy platí: $-1 \le \rho(X,Y) \le 1$.
2. $\mathrm{P}(Y = a + bX) = 1$ práve vtedy, keď $|\rho(X,Y)| = 1$.
3. $\rho(a + bX,\ c + dY) = \rho(X,Y)$, ak súčin $bd > 0$.
4. $\rho(a + bX,\ c + dY) = -\rho(X,Y)$, ak súčin $bd < 0$.
5. Ak X,Y sú nezávislé, potom $\rho(X,Y) = 0$.

Podľa bodu 5 z nezávislosti $X,\ Y$ vyplýva nekorelovanosť (to sme urobili v dôkaze bodu 5 vety 4.2.5). Opačná imlikácia (vo všeobecnosti) neplatí, ako ukazuje nasledujúci príklad.

4.2.18 Príklad. Nech $Y \sim \mathrm{R}\{-1, 0, 1\}$, $X = Y^2$ a uvažujme o náhodnom vektore (X,Y). Čo je možné povedať o závislosti zložiek vektora, o kovariancii a o korelačnom koeficiente?

Riešenie. Aj keď je vektor (X,Y) umelo vytvorený, je to náhodný vektor a ľahko nájdeme tabuľku jeho pravdepodobnostnej funkcie:

$X\backslash Y$	-1	0	1
0	0	1/3	0
1	1/3	0	1/3

Zložky X,Y sú závislé veličiny (sú dokonca funkcionálne závislé, pretože $X = Y^2$). Ukážme, že pritom kovariancia $cov(X,Y) = 0$, a preto aj $\rho(X,Y) = 0$. Kovarianciu môžeme počítať z tabuľky rozdelenia (X,Y) ako $\mathrm{E}(XY)$, pretože $\mathrm{E}(Y) = 0$, (pozri 4.2.12), alebo aj takto:

$$cov(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = \mathrm{E}(Y^2Y) - \mathrm{E}(Y^2)\,\mathrm{E}(Y) = \mathrm{E}(Y^3) - \mathrm{E}(Y^2)\cdot 0 = 0 - 0 = 0,$$

pretože $\mathrm{E}(Y^3) = 0$. Keďže $\rho(X,Y) = 0$, vidíme, že korelačný koeficient *nedetekuje* závislosť medzi X,Y (hoci ide dokonca o funkcionálnu závislosť – teraz kvadratickú).

Táto situácia ilustruje, že nekorelovanosť a nezávislosť nie sú (vo všeobecnosti) totožné pojmy. Zložky X,Y uvažovaného vektora sú nekorelované, ale nie sú nezávislé. Nezávislosť je teda ostrejší vzťah v tom zmysle, že z nezávislosti vyplýva nekorelovanosť (bod 5 vety 4.2.17), ale naopak to (vo všeobecnosti) neplatí.

Ak sa však obmedzíme na prípad vektora (X,Y) s hustotou, ktorého rozdelenie je normálne, tak pojmy *nekorelovanosť* a *nezávislosť* splývajú. To je obsah nasledujúcej vety.

4.2.19 Veta. Nech $(X,Y) \sim \mathrm{N}(a,\ b,\ \sigma_1^2,\ \sigma_2^2,\ \rho)$. Potom $\rho(X,Y) = \rho$, a preto zložky vektora sú nezávislé práve vtedy, keď sú nekorelované.

Dôkaz vety neurobíme, ale obsah musí byť jasný. Symbol $\rho(X,Y)$ znamená Pearsonov korelačný koeficient veličín $X,\ Y$ z definície 4.2.11. Symbol ρ znamená *parameter*, je to piaty parameter rozdelenia (X,Y), t. j. ρ je symbol v zápise $\mathrm{N}(a,\ b,\ \sigma_1^2,\ \sigma_2^2,\ \rho)$. Z článku 3.2 vieme, že *parameter* $\rho = 0$ práve vtedy, keď zložky vektora sú nezávislé (to nie je ťažké ukázať, to sme urobili). Berme ako fakt (bez dôkazu), že pri uvedených predpokladoch, výpočet Pearsonovho korelačného koeficientu $\rho(X,Y)$ vedie k hodnote ρ. Z uvedeného plynie, že v prípade normálneho rozdelenia nezávislosť a nekorelovanosť znamenajú to isté.

Úlohy

4.2.1 Rozdelenie náhodného vektora je dané tabuľkou

X/Y	-2	-1	0	1	2
-1	0.1	0.1	0.05	0	0
0	0	0.15	0.20	0.10	0.05
1	0	0	0.05	0.15	0.05

Uplatnením vety 4.2.1 vypočítajte

a) $\mathrm{E}(X^2 + Y^2)$ b) $\mathrm{E}(X^2Y)$ c) $\mathrm{E}(XY)$ d) $cov(X,Y)$ e) $\rho(X,Y)$

Výsledky a), b), c) overte nájdením rozdelenia $Z = h(X,Y)$ a výpočtom $\mathrm{E}(Z)$ z definície.

4.2.2 Nech $Y = 2 + 3X$, pričom pre X platí: $\mathrm{E}(X) = 1$, $var(X) = 4$. Určte $\mathrm{E}(Y)$, $var(Y)$.

4.2.3 Nech $Y = X(1 + X)$, pričom platí: $\mathrm{E}(X) = 1$, $\mathrm{E}(X^2) = \frac{5}{3}$, $\mathrm{E}(X^3) = 3$, $\mathrm{E}(X^4) = \frac{17}{3}$. Určte $\mathrm{E}(Y)$ a $var(Y)$.

4.2.4 V úlohách 3.1.4, 3.1.5, 3.1.6, 3.1.7, 3.1.8 určte $\mathrm{E}(X)$, $var(X)$, $\mathrm{E}(Y)$, $var(Y)$ a $cov(X,Y)$.

4.2.5 Nech rozdelenie vektora (X,Y) určuje ľavá tabuľka a rozdelenie vektora (U,V) nech určuje pravá tabuľka. Je možné nájsť $a, b \in \mathrm{R}$ tak, aby platilo: $U = aX$, $V = bY$?

$X\backslash Y$	0	1	2	3	4
0	$\frac{3}{24}$	$\frac{2}{24}$	$\frac{1}{24}$	0	0
1	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{2}{24}$	$\frac{1}{24}$	0
2	0	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{2}{24}$	$\frac{1}{24}$
3	0	0	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{3}{24}$

$U\backslash V$	0	2	4	6	8
0	$\frac{3}{24}$	$\frac{2}{24}$	$\frac{1}{24}$	0	0
3	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{2}{24}$	$\frac{1}{24}$	0
6	0	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{2}{24}$	$\frac{1}{24}$
9	0	0	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{3}{24}$

Popíšte, v akom vzťahu sú $cov(X,Y)$ a $cov(U,V)$, resp. v akom vzťahu sú $\rho(X,Y)$ a $\rho(U,V)$.

4.2.6 Nech pravdepodobnostná funkcia vektora (X, Y) je daná tabuľkou

$X \backslash Y$	-1	0	1
-1	0.10	0.20	0.05
0	0.10	0.10	0.10
1	0.10	0.20	0.05

Ukážte, že X, Y sú závislé a pritom nekorelované, naviac, $H(X,Y) = H(X) \times H(Y)$.

4.2.7 Nech $E(X) = -2$, $var(X) = 2$, $E(Y) = 3$, $var(Y) = 4$ a $cov(X, Y) = -1$. Určite

a) $E(2X + 3Y)$, $var(2X + 3Y)$
b) $\rho(2X + 3, 4 - X)$
c) $\rho(2X + 3, 1 - Y)$

4.2.8 Nech $\rho(X, Y) = 0.5$, $var(X) = 1$, $var(Y) = 4$. Určite $var(2X - Y)$.

4.2.9 Nech $X \sim R\{0.2, 0.4, 0.6, 0.8, 1.0\}$ a $Y = 1/X$. Nájdite $cov(X, Y)$ a $\rho(X, Y)$.

4.2.10 Nech pravdepodobnostná funkcia vektora (X, Y) je daná tabuľkou

$X \backslash Y$	0	1	2	3	4
0	0	0	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
1	0	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	0
2	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	0	0

Dokážete určiť znamienko $cov(X, Y)$ bez výpočtu? Môže platiť $var(X + Y) = var(X)$?

4.2.11 Nech pravdepodobnostné funkcie vektorov (X, Y) a (U, V) sú dané tabuľkami

$X \backslash Y$	1	3	5	7
0	.1	.1	0	0
1	.1	.1	.1	0
2	0	.1	.1	.1
3	0	0	.1	.1

$U \backslash V$	1	3	5	7
0	.15	.05	0	0
1	.05	.20	.05	0
2	0	.05	.20	.05
3	0	0	.05	.15

Dokážete bez výpočtu rozhodnúť, ktorý korelačný koeficient má väčšiu hodnotu, $\rho(X, Y)$, či $\rho(U, V)$?

4.2.12 Nech pokus spočíva v 24 hodoch mincou a kockou (súčasne). Nech X predstavuje počet znakov a Y počet šestiek, ktoré padnú v týchto hodoch. Nájdite

a) $E(X + Y)$, $var(X + Y)$,
b) $E(X - Y)$, $var(X - Y)$.

4.2.13 Určte $E(X)$, $var(X)$, $E(Y)$, $var(Y)$ a $cov(X, Y)$, pre náhodný vektor

a) z úlohy 3.1.9,
b) z úlohy 3.1.10.

5 LIMITNÉ VETY

Kým doteraz sme hovorili o jednotlivých náhodných veličinách, resp. o náhodných vektoroch, v tejto kapitole sa budeme zaoberať postupnosťami náhodných veličín (X_n), ktoré sú definované na tom istom pravdepodobnostnom priestore. Pritom z pôvodnej postupnosti (X_n) budeme tvoriť nové postupnosti, a to jednak postupnosť (S_n) čiastočných súčtov

$$S_1 = X_1, \quad S_2 = X_1 + X_2, \quad S_3 = X_1 + X_2 + X_3, \ \dots, \quad S_n = X_1 + X_2 + \dots + X_n, \dots$$

a v článku 5.2 aj postupnosť priemerov

$$X_1, \ \tfrac{1}{2}(X_1 + X_2), \ \tfrac{1}{3}(X_1 + X_2 + X_3), \dots, \ \tfrac{1}{n}(X_1 + X_2 + \dots + X_n), \dots$$

V článku 5.1 pôjde o postupnosť (X_n) nezávislých veličín, ktoré majú alternatívne rozdelenie s tým istým parametrom p, avšak v článku 5.2 budú mať nezávislé veličiny X_n akékoľvek rozdelenie (ale všetky to isté rozdelenie, a pritom také, že má varianciu). To, čo nás bude zaujímať, je limitné rozdelenie súčtov, resp. priemerov. Je zaiste prekvapujúce, že tým limitným rozdelením je v oboch prípadoch Gaussovo normálne rozdelenie. Skutočnosti tejto kapitoly objasňujú centrálnu úlohu normálneho rozdelenia a ako uvidíme, tieto fakty budeme často využívať v štatistike.

5.1 Moivreova-Laplaceova veta

5.1.1 Uvažujme o postupnosti (X_n) nezávislých náhodných veličín, pričom $X_n \sim A(p)$, pre všetky n. Moivreova-Laplaceova veta (voľne povedané) hovorí, že pre veľké n, rozdelenie náhodných veličín S_n je približne normálne rozdelenie. To je zarážajúce napr. aj preto, že veličiny S_n sú diskrétne. Zrejme pre hodnotovú množinu S_n máme

$$H(S_n) = \{0, 1, 2, \dots, n\},$$

pretože S_n je súčtom núl a jednotiek. Nie je problém určiť aj rozdelenie veličiny S_n avšak najprv objasnime, kedy ide o situácie, v ktorých máme do činenia s postupnosťou (X_n), ktorej členy sú nezávislé veličiny s tým istým, alternatívnym rozdelením.

5.1.2 Predstavme si náhodný pokus, ktorý spočíva v n-násobnom nezávislom opakovaní nejakého (jednoduchého) pokusu. V tom jednoduchom pokuse pozorujeme nastatie, resp. nenastatie nejakej náhodnej udalosti A. Tak ako už viackrát predtým (čl. 2.4) s i-tym opakovaním pokusu spojíme predstavu veličiny X_i, ktorá je *indikátorom* udalosti A. To znamená, že X_i nadobúda hodnotu 1 práve vtedy, ak udalosť A v i-tom opakovaní pokusu nastane a X_i nadobúda hodnotu 0 práve vtedy, ak A v i-tom opakovaní pokusu nenastane. Takto máme $X_i \sim A(p)$, kde $p = P(A)$, pre $i = 1, 2, \dots, n, \dots$. Vďaka tomu, že ide o nezávislé opakovanie pokusu, sú veličiny X_i nezávislé.

Pretože $S_n = X_1 + X_2 + \dots + X_n$, S_n predstavuje počet prípadov, keď nastala sledovaná udalosť A, teda S_n predstavuje *absolútnu početnosť* udalosti A v sérii pokusov dĺžky n.

5.1.3 (pokračovanie 5.1.1) Stanovme rozdelenie veličiny S_n. Spomeňme si na záverečný článok kapitoly 1, resp. na odstavce 3.3.4 až 3.3.7, kapitoly 3. Z nich vyplýva, že platí

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

To znamená, že $S_n \sim \mathrm{Bi}(n, p)$, a tak principiálne vieme stanoviť $P(S_n < m)$. Príklad 5.1.4 však ilustruje, že v situáciách, keď je n veľké, sa bez výpočtového prostredia nezaobídeme.

5.1.4 Príklad. Predpokladajme, že riadnou mincou hodíme 1000-krát a chceme určiť pravdepodobnosť toho, že počet padnutých znakov leží v intervale $\langle 490, 510 \rangle$. Podľa 5.1.3 (využitím napríklad MATLABu) máme

$$P(490 \le S_{1000} \le 510) = \sum_{k=490}^{510} P(S_{1000} = k) = \sum_{k=490}^{510} \binom{1000}{k} p^k (1-p)^{1000-k} = 0.49334,$$

pretože $p = 0.5$. Nasledujúca veta umožní približne určiť túto pravdepodobnosť pomocou normálneho rozdelenia.

5.1.5 Veta. (Moivreova–Laplaceova veta) Nech (X_n) je postupnosť nezávislých veličín s alternatívnym rozdelením s parametrom p $(0 < p < 1)$. Potom pre všetky $x \in \mathrm{R}$ platí

$$\lim_{n\to\infty} P\left(\frac{S_n - np}{\sqrt{np(1-p)}} < x \right) = F_N(x),$$

kde F_N je distribučná funkcia normovaného normálneho rozdelenia N(0, 1).

5.1.6 Poznámka. Limitnú vetu 5.1.5 aplikujeme tak, že za aproximáciu neznámej pravdepodobnosti (pre veľké n) vezmeme limitnú hodnotu. Teda pre veľké n kladieme

$$P\left(\frac{S_n - np}{\sqrt{np(1-p)}} < x \right) \approx F_N(x)$$

a ak chceme určiť $P(S_n < k)$, tak

$$P(S_n < k) = P(S_n - np < k - np) = P\left(\frac{S_n - np}{\sqrt{np(1-p)}} < \frac{k - np}{\sqrt{np(1-p)}} \right) \approx F_N\left(\frac{k-np}{\sqrt{np(1-p)}} \right).$$

5.1.7 Príklad. Vypočítajme príklad 5.1.4 použitím vety 5.1.5 a poznámky 5.1.6. Zrejme pre udalosti platí

$$\{S_n < 490\} \cup \{490 \le S_n < 511\} = \{S_n < 511\},$$

a preto

$$P(490 \le S_n \le 510) = P(490 \le S_n < 511) = P(S_n < 511) - P(S_n < 490) \approx$$

$$\approx F_N\left(\frac{511 - np}{\sqrt{np(1-p)}}\right) - F_N\left(\frac{490 - np}{\sqrt{np(1-p)}}\right) = F_N\left(\frac{511 - 500}{15.81}\right) - F_N\left(\frac{490 - 500}{15.81}\right) = F_N(0.70) - F_N(-0.63) =$$

$$= 0.75804 - 0.26435 = 0.49369.$$

Presnou hodnotou (zaokrúhlenou na 5 desatinných miest) je 0.49334. Vidíme, že absolútna chyba aproximácie je menšia ako 4 desaťtisíciny a môžeme byť spokojní. Pretože binomické rozdelenie (teda diskrétne rozdelenie) aproximujeme normálnym (teda spojitým rozdelením) niekedy sa používa korekcia, ktorou dosiahneme ešte presnejšiu aproximáciu, čo ilustruje príklad 5.1.8.

5.1.8 Príklad. Riešme príklad 5.1.4 použitím vety 5.1.5, poznámky 5.1.6 a uplatnením korekcie. Tá spočíva v tom, že kladieme

$$P(490 \le S_n \le 510) \approx F_N\left(\frac{510.5 - np}{\sqrt{np(1-p)}}\right) - F_N\left(\frac{489.5 - np}{\sqrt{np(1-p)}}\right) = F_N\left(\frac{510.5 - 500}{15.81}\right) - F_N\left(\frac{489.5 - 500}{15.81}\right) =$$

$$= F_N(0.664) - F_N(-0.664) = 2F_N(0.664) - 1 = 0.49331.$$

Vidíme, že teraz je aproximácia rádovo lepšia (ako bola bez korekcie). Vo všeobecnosti sa nedá povedať, aké n je dostatočne veľké, aby aproximácia dávala dobré výsledky. Zistilo sa však, že ak platí: $np(1-p) > 9$, tak výsledky sú uspokojivé.

5.1.9 Príklad. Počas skúšky spoľahlivosti dochádza k zničeniu výrobku s pravdepodobnosťou $p = 0.02$. Nájdime pravdepodobnosť toho, že v sérii 500 výrobkov je počet zničených

a) menší ako 12,

b) menší ako 12, a pritom viac ako 6.

Riešenie. a) Najprv bez korekcie:

$$P(S_{500} < 12) = P\left(\frac{S_{500} - 500\ 0.02}{\sqrt{500\ 0.02\ 0.98}} < \frac{12 - 500\ 0.02}{\sqrt{500\ 0.02\ 0.98}} \right) \approx F_N\left(\frac{12 - 500\ 0.02}{\sqrt{500\ 0.02\ 0.98}} \right) = F_N(0.64) = 0.73891.$$

Teraz s korekciou:

$$P(S_{500} < 12) = P(S_{500} \le 11) \approx F_N\left(\frac{11.5 - 500\ 0.02}{\sqrt{500\ 0.02\ 0.98}} \right) = F_N(0.48) = 0.68439.$$

Presná hodnota zaokrúhlená na 5 miest sa rovná 0.69794 (je to hodnota distribučnej funkcie rozdelenia $\mathrm{Bi}(n, p)$ pre $n = 500$, $p = 0.02$ v bode $x = 12$). Časť (b) príkladu je úloha 5.1.1.

5.1.10 Príklad. Ukážme, že ak (X_n) je postupnosť nezávislých veličín s alternatívnym rozdelením s parametrom p, tak pre veľké n

$$S_n \text{ má približne rozdelenie } \mathrm{N}(np,\ np(1-p)).$$

Riešenie. Pozrime sa najprv na vetu 5.1.5 a uvedomme si, že náhodná veličina, ktorá vystupuje na ľavej strane vzťahu je normovaná veličina S_n, pretože

$$E(S_n) = E(X_1 + X_2 + \ldots + X_n) = E(X_1) + E(X_2) + \ldots + E(X_n) = np,$$
$$var(S_n) = var(X_1 + X_2 + \ldots + X_n) = var(X_1) + var(X_2) + \ldots + var(X_n) = np(1-p).$$

Podľa vety 5.1.5 pre veľké n má veličina $\frac{S_n - np}{\sqrt{np(1-p)}}$ približne normálne rozdelenie N(0, 1), odkiaľ približne

$$S_n - np \sim \mathrm{N}(0,\ np(1-p)) \text{ a nakoniec, } S_n \sim \mathrm{N}(np,\ np(1-p)).$$

Úlohy

5.1.1 Nájdite výsledok pre časť (b) príkladu 5.1.9.

5.1.2 Pravdepodobnosť výroby nepodarku $p = 0.03$. Uvažujme o počte nepodarkov v sérii 400 výrobkov. S akou pravdepodobnosťou počet nepodarkov bude v intervale $\langle 8, 15 \rangle$?

5.1.3 Hodíme riadnou hracou kockou 150-krát. S akou pravdepodobnosťou padne šestka aspoň 20-krát?

5.1.4 Hádžeme riadnou hracou kockou dovtedy, kým nepadne šestka 20-krát. S akou pravdepodobnosťou budeme hádzať aspoň 151-krát?

5.1.5 Pravdepodobnosť udalosti A v pokuse sa rovná 0.6. Určte pravdepodobnosť toho, že v sérii 200 opakovaní pokusu sa udalosť A objaví aspoň 111-krát, a pritom menej ako 130-krát.

5.2 Centrálna limitná veta – CLV

5.2.1 Voľne (a zatiaľ nepresne) povedané, CLV hovorí, že ak (X_n) je postupnosť nezávislých, rovnako rozdelených náhodných veličín, tak rozdelenie súčtu $S_n = X_1 + X_2 + \ldots + X_n$ (pre dostatočne veľké n) je temer rozdelenie normálne. Predtým než sformulujeme CLV ako matematickú vetu, pouvažujme nad hore uvedenou voľnou formuláciou. Je na mieste pýtať sa, kedy máme do činenia s postupnosťou nezávislých, rovnako rozdelených náhodných veličín. Inými slovami: Aké situácie modelujeme takými postupnosťami?

5.2.2 Uvažujme o nejakom (jednoduchom) pokuse, v ktorom pozorujeme náhodnú veličinu X. Nech jej rozdelenie určuje funkcia f. Teraz nezávisle (a za rovnakých podmienok) opakujeme tento pokus nielen konečne veľakrát, ako tomu bolo v závere kapitoly 1, ale potenciálne vzaté, nekonečne veľa ráz. Veličinu X v i-tom opakovaní označme symbolom X_i. Takto vzniká postupnosť (X_i) náhodných veličín, o ktorej môžeme povedať toto:

Kým veličina X je veličinou v jednoduchom pokuse, veličiny X_i sú veličinami v novom (v zloženom) pokuse, ktorého pravdepodobnostný priestor je komplikovanejší a nepokúšame sa o ňom hovoriť. To, čo je pre nás dôležité, je chápať, že veličiny X_i majú rovnaké rozdelenie. Ich rozdelenie je totožné s rozdelením, ktoré má skúmaná veličina X a je teda určené funkciou f. To preto, že ide o opakovanie toho istého pokusu za rovnakých podmienok. Veličiny X_i sú pritom, samozrejme, rôzne veličiny, dokonca (a to je podstatné) sú to nezávislé veličiny. Veď ten jednoduchý pokus sme opakovali takým spôsobom, aby výsledok i-tého opakovania neovplyvňoval výsledok iného opakovania, teda išlo o nezávislé opakovania. Je kľúčové (zvlášť pre štatistiku) pochopiť túto myšlienkovú konštrukciu. V reálnych situáciách budeme mať do činenia vždy s konečnou postupnosťou n členov, kde však n je niekedy dosť veľké prirodzené číslo.

5.2.3 Príklad. Ilustrujme obsah odseku 5.2.2. Uvažujme o pokuse, ktorý spočíva v opakovanom hode hracou kockou. Jednoduchý pokus je (jeden) hod kockou. Skúmaná veličina X nech je počet bodov, ktoré padnú na kocke. Opakujme hod napr. 50-krát. S i-tým hodom spojme predstavu veličiny X_i. Takto máme n-ticu, resp. 50-ticu, $X_1, X_2, \ldots, X_{50}$, nezávislých veličín, ktoré majú rovnaké rozdelenie. Ak je kocka riadna, ide o rovnomerné rozdelenie R{1/6, 1/6,..., 1/6}; ak je kocka falošná, rozdelenie veličín X_i určuje vektor $(p_1, p_2, \ldots, p_6)$, v ktorom aspoň dve zložky sa nerovnajú číslu 1/6. Teraz si predstavme, že počet hodov n konverguje do nekonečna. Touto úvahou dostávame postupnosť $X_1, X_2, \ldots, X_n, \ldots$ nezávislých, rovnako rozdelených veličín.

Teraz uvažujme o postupnosti (S_n) čiastočných súčtov. Z odseku 3.3.6 vieme, že rozdelenia súčtov $S_2 = X_1 + X_2$, $S_3 = X_1 + X_2 + X_3$, $S_4 = X_1 + X_2 + X_3 + X_4$, už nie sú rovnomerné rozdelenia. S rastúcim počtom sčítancov rozdelenia získavajú zvonovitý tvar (obr. 3-1, str. 90). Podľa CLV platí, že keď počet sčítancov konverguje do nekonečna, rozdelenie normovanej veličiny $S_n = X_1 + X_2 + \ldots + X_n$ sa blíži k normálnemu rozdeleniu (a to bez ohľadu na to, či je kocka normálna, alebo falošná). Uvedomme si, že v tomto prípade sú veličiny X_i diskrétne, a preto pre každé n aj súčet S_n je diskrétna veličina, napr. $H(S_{50}) = \{50, 51, \ldots, 300\}$. Ak však n je dostatočne veľké, rozdelenie súčtu $S_n = X_1 + X_2 + \ldots + X_n$ môžeme aproximovať rozdelením normálnym – teda spojitým rozdelením. Toto ilustrujú úlohy 5.2.2 a 5.2.3.

5.2.4 Poznámka. Nech (X_n) je postupnosť nezávislých, rovnako rozdelených náhodných veličín, pričom $\mathrm{E}(X_n) = m$ a $var(X_n) = b^2$ (žiadne konkrétne rozdelenie nemáme na mysli). Aj keď rozdelenie veličiny S_n nepoznáme (samozrejme, veď nepoznáme rozdelenie X_i), vieme určiť strednú hodnotu a varianciu veličiny S_n. Zrejme platí:

$$\mathrm{E}(S_n) = \mathrm{E}(X_1 + X_2 + \ldots + X_n) = \mathrm{E}(X_1) + \mathrm{E}(X_2) + \ldots + \mathrm{E}(X_n) = nm,$$

$$var(S_n) = var(X_1 + X_2 + \ldots + X_n) = var(X_1) + var(X_2) + \ldots + var(X_n) = nb^2.$$

V CLV bude vystupovať náhodná veličina, ktorá vzniká normovaním S_n (odsek 4.2.7), teda od S_n odčítame jej strednú hodnotu a predelíme smerodajnou odchýlkou $\sqrt{var(S_n)} = b\sqrt{n}$. Normovaním S_n dostávame veličinu

$$\frac{S_n - nm}{b\sqrt{n}}.$$

5.2.5 Centrálna limitná veta. Nech (X_n) je postupnosť nezávislých, rovnako rozdelených náhodných veličín, pričom to rozdelenie má strednú hodnotu m a kladnú varianciu b^2. Potom pre každé $x \in \mathrm{R}$ platí

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{S_n - nm}{b\sqrt{n}} < x\right) = F_N(x),$$

teda limitné rozdelenie veličiny $\frac{S_n - nm}{b\sqrt{n}}$ je normované normálne rozdelenie N(0, 1).

5.2.6 Poznámka. Centrálna limitná veta (CLV) nemôže hovoriť o limitnom rozdelení samotných súčtov S_n, pretože pre $n \to \infty$ stredná hodnota (a aj variancia) S_n rastú nad všetky medze. CLV aplikujeme tak, že pre veľké n považujeme rozdelenie veličiny $\frac{S_n - nm}{b\sqrt{n}}$ za približne normálne rozdelenie N(0, 1). To znamená, že pre veľké n

$\frac{S_n - nm}{b\sqrt{n}}$ má približne rozdelenie N(0, 1),

odkiaľ

$S_n - nm$ má približne rozdelenie $\mathrm{N}(0, nb^2)$,

a preto

S_n má približne rozdelenie $\mathrm{N}(nm, nb^2)$

a priemer

$\frac{1}{n} S_n$ má približne rozdelenie $\mathrm{N}(m, \frac{b^2}{n})$.

Je celkom prirodzené pýtať sa, aké n môžeme považovať za (dostatočne) veľké. Nie je možné dať jednoznačnú odpoveď, pretože odpoveď závisí na rozdelení samotných veličín X_i, z ktorých vytvárame S_n. Napr. ak $X_i \sim \mathrm{R}(0, 1)$, tak už pre $n = 12$ dostávame neočakávane dobrú zhodu medzi rozdelením $\frac{S_n - nm}{b\sqrt{n}}$ a rozdelením N(0, 1) (pozri úlohu 5.2.1).

5.2.7 Príklad. Uvažujme o životnosti žiariviek istého typu. Predpokladajme, že životnosť je veličina s rozdelením Exp(0.01), t. j. očakávaná životnosť je 100 hodín. Predstavme si, že nepoznáme obsah vety 3.3.7 o tom, že súčet 20-tich nezávislých veličín s rozdelením Exp(0.01) je veličina s rozdelením Erl(20, 0.01), a preto chceme aplikovať CLV na výpočet pravdepodobnosti toho, že tých 20 žiariviek (zapájame novú, keď predchádzajúca vyhorela) vydrží aspoň 1500 hodín. To znamená, že chceme aproximovať

$$P(S_{20} > 1500),$$

keď $S_{20} = X_1 + X_2 + \ldots + X_{20}$, pričom X_i sú nezávislé a $X_i \sim \mathrm{Exp}(0.01)$, pre $i = 1, 2, \ldots, 20$.

Riešenie. $X_i \sim \mathrm{Exp}(0.01)$, pre $i = 1, 2, \ldots, 20$, a preto $m = E(X_i) = 100$, $b^2 = var(X_i) = 10\,000$. Podľa 5.2.6 veličina S_{20} má približne rozdelenie $N(nm, nb^2)$, t. j. N(2000, 200 000). Máme

$$P(S_{20} > 1500) = 1 - P(S_{20} < 1500) = 1 - P\left(\frac{S_{20} - 2000}{\sqrt{200000}} < \frac{1500 - 2000}{\sqrt{200000}}\right) = 1 - F_N(-1.12) = 0.86864.$$

Pre porovnanie, presná hodnota (zaokrúhlená na 5 desatinných miest), ktorú sme získali integrovaním hustoty Erlangovho rozdelenia sa rovná 0.87522. Vidíme, že CLV poskytla veľmi dobrú aproximáciu hľadanej pravdepodobnosti (chyba je menšia ako 0.007).

Úlohy

5.2.1 Nech (X_i) je postupnosť nezávislých náhodných veličín s rozdelením R(0, 1). V odseku 5.2.6 sme poznamenali, že už pre $n = 12$, má veličina $\frac{S_n - nm}{b\sqrt{n}}$ približne normálne rozdelenie. Pretože pre rozdelenie R(0, 1) máme $m = E(X) = \frac{1}{2}$ a $var(X) = b^2 = \frac{1}{12}$, tak pre $n = 12$ veličina $\frac{S_n - nm}{b\sqrt{n}}$ má tvar $S_{12} - 6$. Teda $S_{12} - 6$ má približne normálne rozdelenie N(0, 1). Pre $n = 30$ dosiahneme lepšiu aproximáciu. Aký tvar má veličina $\frac{S_n - nm}{b\sqrt{n}}$ v prípade $n = 30$?

5.2.2 Riadnou hracou kockou hodíme 36-krát. Určte (resp. aproximujte) pravdepodobnosť toho, že súčet bodov, ktoré padnú v pokuse, bude aspoň 144. Nájdite chybu v úvahe:

Pretože 144/36 = 4, pričom na kocke padnú aspoň 4 body s pravdepodobnosťou 0.5, tak (vďaka nezávislosti) hľadaná pravdepodobnosť sa rovná $(0.5)^{36} = 1.46\ 10^{-11}$.

5.2.3 Teraz uvažujme o (hypoteticky) falošnej kocke, na ktorej jednotlivé steny s bodmi 1, 2, 3, ..., 6 padajú s pravdepodobnosťami: (0.1, 0.1, 0.1, 0.1, 0.2, 0.4). Predstavme si, že takouto kockou hodíme 36-krát. Určte (resp. aproximujte) pravdepodobnosť toho, že súčet bodov, ktoré padnú v pokuse, bude aspoň 144.

5.2.4 Životnosť batérie istého typu je veličina so strednou hodnotou 55 (hod.) a smerodajnou odchýlkou 8 (hod.). Keď batéria dožije, je vzápätí nahradená novou rovnakého typu. Aproximujte pravdepodobnosť toho, že 20 batérií nám postačí na aspoň 1000 hodín práce.

5.2.5 Nech (X_i) je postupnosť nezávislých veličín s rozdelením N(0, 1). Vieme, že veličiny X_i^2 už nemajú normálne rozdelenie (pozri 2.5.11). Zdôvodnite, prečo náhodné veličiny

$$S_n = X_1^2 + X_2^2 + \cdots + X_n^2$$

majú pre veľké n približne normálne rozdelenie. Otázka znie: Aké sú parametre toho normálneho rozdelenia? (Využite fakt, ktorý sme nedokazovali: Ak $X_i \sim N(0, 1)$, tak $E(X_i^4) = 3$).

6 ZÁKLADNÉ POJMY ŠTATISTIKY

6.1 Od pravdepodobnosti k štatistike

6.1.1 Začnime niekoľkými myšlienkami o tom, čo je štatistika ako vedná disciplína. Zjednodušene povedané, cieľom štatistiky je získavať zmysluplnú informáciu z daného súboru dát. To zjednodušenie je v slovách „... z daného súboru dát". Dáta by nemali „byť dané", ale mali by byť premyslene získavané tak, aby čo najväčšou mierou prispeli k riešeniu konkrétnej úlohy. Úloha štatistika by mala začať už vtedy, keď formulujeme cieľ nášho snaženia a uvažujeme o tom, *ako* budeme dáta získavať, alebo napr. *koľko* ich potrebujeme získať, aby cieľ úlohy bol dosiahnutý.

Základnou črtou situácií, v ktorých využívame štatistické metódy, je prítomnosť variability. Variabilitu (premenlivosť), chápeme ako niečo, čo sa prejavuje tým, že údaje, teda dáta, sú viac-menej rôzne a na pohľad náhodné. Získavame ich meraním, pozorovaním, sumarizovaním odpovedí na otázky v dotazníkoch atď. V ďalšom texte budeme na tieto akcie používať slovo „meranie" a treba ho chápať v naznačenom širokom zmysle.

Ako v každom vednom odbore, aj v štatistike si kladieme otázky o svete okolo nás a hľadáme odpovede. Pre štatistiku je typické práve to, že sa zapodieva takými problémami, v ktorých (v nejakej podobe) vystupuje variabilita. Variabilitu merania, resp. variabilitu nejakého procesu, považujeme za (neodstrániteľnú) vlastnosť reality. Spomeňme aspoň niektoré zdroje tohoto javu:

- Podmienky, za ktorých sa dáta získavajú, sa v čase menia a nie je vždy možné držať ich pod kontrolou, t. j. zaručiť ich nemennosť. Tak je to napr. vo výrobných procesoch, alebo pri meraniach vykonávaných na živých organizmoch.
- Samotný proces merania (akokoľvek presného) podlieha okolnostiam, vďaka ktorým opakované merania tej istej veličiny nedávajú úplne rovnaký výsledok.
  Príklad: každé (viac-menej) presné meranie.
- Veľmi často nie je možné realizovať meranie na všetkých objektoch uvažovaného *základného súboru* (tento pojem neskôr objasníme), a preto meriame len na prvkoch „výberu". Na jeho základe urobíme záver o celku (tejto myšlienkovej konštrukcii hovoríme *štatistická indukcia*). Záver (o celku) je takto závislý od konkrétneho výberu a tam – v okamihu realizovania výberu – bola prítomná „pani Náhoda".

Zhrňme: Štatistika sa zaoberá zmysluplným získavaním dát a ich analýzou, s cieľom prispieť k riešeniu konkrétneho problému ľudskej praxe. Základnou črtou situácií, v ktorých sa využívajú štatistické metódy je to, že v nich je (v nejakej podobe) prítomná variabilita. Ak na prvkoch výberu sledujeme jeden znak, chápeme (modelujeme) ho ako náhodnú veličinu. Konkrétne zistené hodnoty znaku považujeme za realizácie tej náhodnej veličiny. Ak na prvkoch výberu pozorujeme súčasne dva, alebo viac znakov, tak ide o pozorovanie hodnôt náhodného vektora (dvoj, resp. viacrozmerného).

6.1.2 V predchádzajúcom odseku sme použili pravdepodobnostné pojmy: náhodná veličina, resp. náhodný vektor. Štatistika totiž na modelovanie situácií (v ktorých má nájsť odpovede na položené otázky) používa pravdepodobnostné pojmy, a tak je niekedy rozdiel medzi úlohou z pravdepodobnosti a úlohou zo štatistiky naozaj jemný. Preto vezmime situácie, v ktorých formulujeme na jednej strane úlohu typicky pravdepodobnostnú (odseky 6.1.3, 6.1.5) a na druhej strane úlohu typicky štatistickú (odseky 6.1.4, 6.1.6).

6.1.3 Príklad. (Tombola I) Predstavme si, že na spoločenskej udalosti sme organizátori tomboly, a tak sme si prichystali dostatočný počet očíslovaných kupónov a snažili sa ich predať čo najviac. Nakoniec sa nám podarilo predať 30 kupónov (tých 30 je očíslovaných od 1 do 30). Keď v programe príde na losovanie cien tomboly (povedzme, že máme 5 cien), z osudia, do ktorého sme uložili lístky s číslami od 1 do 30, budeme náhodne postupne ťahať výhercu piatej ceny, potom štvrtej ceny atď., a nakoniec výhercu prvej ceny (výberom bez vrátenia). Typickými pravdepodobnostnými úlohami sú:

a) Aká je pravdepodobnosť toho, že čísla na 5-tich vytiahnutých lístkoch vytvorili klesajúcu postupnosť čísel?
b) Aká je pravdepodobnosť toho, že všetky vytiahnuté čísla sú menšie ako 16?

Riešenie. (a) Vzhľadom na formuláciu otázky, uvažujme pravdepodobnostný priestor všetkých usporiadaných 5-tíc čísel od 1 po 30. Zrejme počet všetkých možných výsledkov sa rovná V(30, 5). Keďže počet všetkých klesajúcich 5-tíc sa rovná C(30, 5), hľadaná pravdepodobnosť sa rovná podielu C(30, 5) / V(30, 5) = 0.0083.

(b) Teraz môžeme výsledok pokusu chápať ako neusporiadanú 5-ticu rôznych čísel. Počet všetkých možných výsledkov sa rovná C(30, 5) a počet tých, ktoré sú priaznivé pre uvažovanú udalosť, sa rovná C(15, 5). Preto výsledkom je podiel C(15, 5) / C(30, 5) = 0.0211.

Pre pravdepodobnostnú úlohu je charakteristické to, že poznáme pomery v osudí (vieme koľko a akých lístkov obsahuje) a máme stanoviť pravdepodobnosť toho, že náhodný výber dopadne takým, resp. onakým spôsobom (t. j. že ako výsledok zistíme napr. klesajúcu postupnosť, alebo 5-ticu, ktorej všetky zložky sú menšie ako 16).

6.1.4 Príklad. (Tombola II) Predstavme si, že sme účastníkmi večierka a ako pozorovatelia si všimneme, že vyhrávajúce čísla v tombole sú čísla $x_1 = 11$, $x_2 = 7$, $x_3 = 15$, $x_4 = 8$, $x_5 = 4$. Napadne nás otázka: Koľko kupónov sa organizátorom tomboly podarilo predať? Inými slovami: Koľko lístkov bolo vložených do osudia?

Teraz ide o úlohu v istom zmysle obrátenú. Poznáme výsledok náhodného pokusu (máme dáta) a z neho (resp. z nich, zo zistených $x_1, x_2, \dots, x_5$) chceme odhadnúť počet lístkov v osudí. Ide o štatistickú úlohu. Je zrejmé, že tam muselo byť aspoň 15 lístkov. Zároveň už vieme, že je dosť nepravdepodobné, že ich tam bolo až 30, a to vďaka 6.1.3, úloha (b).

6.1.5 Príklad. Predpokladajme, že technológia zaručuje, že životnosť výrobku X (v hod.) sa riadi exponenciálnym rozdelením, $X \sim \mathrm{Exp}(\lambda)$, kde $\lambda = 0.001$. Predpokladajme, že z uvažovanej produkcie vyberieme (výberom bez vrátenia) náhodne 10 výrobkov. To znamená, že 10-krát zopakujeme náhodný pokus – náhodné vybratie výrobku. Celková produkcia je však v porovnaní s rozsahom $n = 10$ taká veľká, že ak sme napr. už 4 výrobky vybrali, pomery v celkovej produkcii sa prakticky nezmenili (piaty výrobok vyberáme za rovnakých okolností ako prvý). *Preto sa na výber bez vrátenia môžeme pozerať ako na výber s vrátením.* Výber chápeme ako 10-násobné *nezávislé opakovanie* toho istého pokusu – tým pokusom je náhodný výber výrobku z celkovej produkcie. Určíme pravdepodobnosť toho, že

a) životnosť každého vybratého bude aspoň 300 hod.
b) životnosť aspoň jedného vybratého prevýši 1500 hod.

Riešenie. Toto je typická pravdepodobnostná úloha, keď predpokladáme, že rozdelenie životnosti X je exponenciálne rozdelenie (so známym parametrom). Analogickú situáciu sme rozoberali v kapitole 3 (pozri 3.2.4 a 3.2.5). V čl. 6.2 vysvetlíme, že náhodný výber budeme modelovať 10-ticou $X_1, X_2, \dots, X_{10}$ nezávislých náhodných veličín X_i, pre ktoré – podľa predpokladu – platí: $X_i \sim \mathrm{Exp}(\lambda)$, kde $\lambda = 0.001$. Teraz chceme určiť

a) $P(\bigcap_{i=1}^{10}(X_i \geq 300))$, t. j. P(životnosť každého vybratého bude aspoň 300 hod.)
b) $P(\bigcup_{i=1}^{10}(X_i \geq 1500))$, t. j. P(životnosť aspoň jedného vybratého prevýši 1500 hod.)

Zrejme $P(\bigcap_{i=1}^{10}(X_i \geq 300)) = \prod_{i=1}^{10} P(X_i \geq 300) = \prod_{i=1}^{10} e^{-0.001 * 300} = (e^{-0.3})^{10} = 0.050$,

resp. $P(\bigcup_{i=1}^{10}(X_i \geq 1500)) = 1 - P(\bigcap_{i=1}^{10}(X_i < 1500)) = 1 - \prod_{i=1}^{10}(1 - e^{-1.5}) = 0.920$.

6.1.6 Príklad. Sformulujme teraz štatistickú úlohu, ktorú môžeme považovať za protipól pravdepodobnostnej úlohy z predchádzajúceho odseku. Predpokladajme, že konkrétnou technológiou vyrábané výrobky majú životnosť, ktorú modelujeme exponenciálnym rozdelením, avšak teraz parameter rozdelenia λ *nie je známy*. Z celkovej produkcie náhodne vyberieme 10 výrobkov a na základe merania ich životností, t. j. na základe zistených (nameraných) hodnôt $x_1 = 855$, $x_2 = 1048$, …, $x_{10} = 792$, chceme odhadnúť napr. strednú hodnotu životnosti, resp. samotný parameter λ. Ide o situáciu, v ktorej *pravdepodobnostný model* predstavuje exponenciálne rozdelenie a neznámy parameter toho rozdelenia chceme odhadnúť na základe dát, ktoré poskytol experiment (riešenie situácie ukážeme v kapitole o bodových odhadoch).

6.2 Štatistický súbor a náhodný výber

6.2.1 Jedným zo základných pojmov štatistiky je *štatistický súbor* (tiež *populácia*). Je to konečná množina, na prvkoch ktorej má zmysel pozorovať určitý znak, t. j. prítomnosť nejakej vlastnosti, či hodnotu nejakej veličiny. Cieľom štatistiky je prispieť k zodpovedaniu nejakej otázky, ktorá sa týka štatistického súboru (často sa namiesto štatistický súbor hovorí *základný súbor*). Štatistika robí záver o základnom súbore prostredníctvom analyzovania údajov zistených na *niektorých* jednotkách, t. j. na niektorých prvkoch, základného súboru (zopakujme, že prenosu poznania, ktoré sme získali z jednotiek výberu, na celok, sa hovorí *štatistická indukcia*). Jednotky základného súboru môžu byť veci, ľudia, či iné živé organizmy. Skrátka všeličo. Avšak zrejme nie je možné povedať, že akákoľvek konečná množina predstavuje štatistický súbor – uvažujte prečo! (Pomôcka: prečítajte si pozorne druhú vetu tohoto odseku, resp. 6.2.2.) V príklade 6.1.6 štatistický súbor tvorí množina výrobkov vyrábaných uvažovanou technológiou. V tomto prípade nie je možné presne špecifikovať všetky prvky štatistického súboru, pretože súbor, ako taký, sa v čase neustále mení. Hovoríme, že ide o *hypotetický* základný súbor. V takomto prípade parametre pravdepodobnostného modelu vhodne modelujú (opisujú) vlastnosti samotného technologického postupu. A to je dôvod, prečo chceme – na základe dát – tie parametre odhadnúť.

6.2.2 Podľa odseku 6.2.1 každú konečnú množinu nepovažujeme za štatistický súbor. Ak množina má byť štatistickým súborom, na jej prvkoch musí mať zmysel pozorovať nejaký znak, ktorý nesie informáciu o konkrétnej jednotke (o konkrétnom prvku) základného súboru. Naviac, základný súbor je síce množina konečná, avšak počet jej prvkov býva taký veľký, že obyčajne nie je možné pozorovať (merať) znak na všetkých jeho prvkoch. Dôvody sú zrejmé: časová, resp. finančná náročnosť, alebo dokonca deštrukcia jednotky počas merania, ako napr. pri meraní životnosti výrobku (žiaroviek, batérií a pod.). Preto temer vždy pracujeme s *výberom* zo základného súboru. To znamená, že preskúmame, resp. zisťujeme hodnotu znaku, len na jednotkách *výberu* (*výberového súboru*). Počet jednotiek, ktoré sa dostali do výberu, sa nazýva *rozsah* výberu. Kľúčovou vlastnosťou výberu je jeho „reprezentatívnosť". Je pochopiteľné, že chceme, aby výber bol verným obrazom celku, t. j. základného súboru, pretože zistenia, ktoré máme z výberu, chceme preniesť, zovšeobecniť, na základný súbor. Ako to dosiahnuť? Ako realizovať reprezentatívny výber, povedzme, rozsahu n zo základného súboru, ktorý má N jednotiek? V rámci nasledujúceho príkladu ukážeme náhodný mechanizmus, ktorý zaručuje, že každá n-tica jednotiek základného súboru má rovnakú pravdepodobnosť vytvoriť výberový súbor. Takému náhodnému výberu hovoríme *jednoduchý náhodný výber*.

6.2.3 Príklad. Nech základným súborom je množina všetkých študentov prvého ročníka našej fakulty (pre konkrétnosť, nech napr. počet prvákov $N = 400$). Predpokladajme, že nás zaujíma rozloženie IQ na uvažovanom základnom súbore. Pretože merať IQ na všetkých jednotkách základného súboru by bolo (z viacerých hľadísk) náročné, chceme zistiť hodnoty IQ len na prvkoch náhodného výberu, povedzme, rozsahu 40. Navrhnime postup, ktorým by sme získali jednoduchý náhodný výber rozsahu $n = 40$.

Riešenie. Každý študent prvého ročníka napíše na lístok svoje meno a k nemu pre istotu aj rodné číslo. Do škatule dáme všetkých 400 lístkov, zatrasieme ňou a vyberieme náhodne jeden po druhom 40 lístkov (pre dôkladnosť, po každom výbere škatuľou dobre zatrasieme). Realizácia výberu končí vybratím štyridsiateho lístka. Na stole máme 40 rôznych lístkov, pretože vybraté lístky sme (pred ťahaním ďalšieho) späť do škatule nedávali. Intuitívne je asi jasné, že sme naplnili obsah požadovaného, pretože nevidíme dôvod, prečo by nejakých konkrétnych 40 lístkov malo väčšiu, alebo menšiu šancu tvoriť výber. Ak situáciu modelujeme tak, ako v článku 1.2, tak môžeme určiť pravdepodobnosť p toho, že výber tvorí konkrétnych 40 lístkov: $p = 1/C(400, 40) \approx 5.10^{-56}$.

6.2.4 Poznámka. Zmyslom nasledujúceho príkladu 6.2.5 je spresniť predstavu o pojmoch základný súbor, rozloženie znaku na základnom súbore, náhodný výber a o tom, ako budeme náhodný výber modelovať (t. j. ako s ním budeme pracovať). Z toho, čo sme povedali v odsekoch 6.2.1 a 6.2.2, je náhodný výber akousi vzorkou údajov, ktoré sme získali na vybratých jednotkách základného súboru. Aby sme však mohli posúdiť rôzne postupy spracovania získaných dát (pozri odsek 6.2.8), je nutné náhodný výber teoreticky modelovať. Pochopiť, že za model náhodného výberu vezmeme n-ticu veličín $X_1, X_2, \dots, X_n$ (!?) s nejakými vlastnosťami, naozaj nie je jednoduché. Je to však kľúčové k pochopeniu základných myšlienok štatistiky.

6.2.5 Príklad. Začnime s pravdepodobnostnou úlohou. Predpokladajme, že výrobky sú vyrábané tak, že pravdepodobnosť výroby chybného sa rovná 0.06. Aká je pravdepodobnosť toho, že v sérii 100 výrobkov bude chybných menej ako 8? V článku 5.1 sme situáciu (zatiaľ pravdepodobnostnú situáciu) modelovali takto:

X_i nech je indikátor toho, či je i-tý výrobok nepodarok, to znamená, že

$X_i = 1$, ak i-tý výrobok je nepodarok, resp. $X_i = 0$, ak i-tý výrobok je dobrý výrobok.

Podľa predpokladu, $X_i \sim \mathrm{A}(p)$, $p = 0.06$. To, čo nás zaujíma je $\mathrm{P}(X_1 + X_2 + \dots + X_{100} < 8)$, lebo veličina $X_1 + X_2 + \dots + X_{100}$ je vlastne *počítadlo* nepodarkov v uvažovanej sérii. Ide o typickú pravdepodobnostnú úlohu. V odsekoch 5.1.2 a 5.1.3 sme zistili, že ak môžeme predpokladať, že X_i sú nezávislé, tak počet nepodarkov v sérii, teda súčet $X_1 + X_2 + \dots + X_{100}$, je veličinou s binomickým rozdelením Bi(100, 0.06).

Výpočet pravdepodobnosti $\mathrm{P}(X_1 + X_2 + \dots + X_{100} < 8)$ môžeme urobiť použitím nejakého výpočtového prostredia, alebo hľadanú pravdepodobnosť môžeme aproximovať tak, ako o tom hovorí Laplaceova-Moivreova veta (odsek 5.1.5).

Je prirodzené zamyslieť sa, na základe čoho v reálnej situácii môžeme predpokladať, že $X_i \sim \mathrm{A}(p)$, že $p = 0.06$, resp., že X_i sú nezávislé. Jasne pomenujme o čom hovoríme:

a) To, že rozdelenie veličín X_i je *alternatívne* rozdelenie, je výsledok pravdepodobnostnej úvahy (pravdepodobnostného modelovania). To sme si osvojili už v kapitole 2.
b) To, že X_i môžeme považovať za nezávislé, je treba kriticky diskutovať (čo hneď urobíme). Kľúčovým je aj fakt, že všetky X_i majú rozdelenie $\mathrm{A}(p)$ *s tou istou hodnotou* parametra p (čo znamená, že všetky X_i majú rovnaké rozdelenie).
c) Ako sa v reálnej situácii dostaneme k hodnote parametra p (= pravdepodobnosť výroby nepodarku), je výsledok štatistických postupov, o ktorých budeme hovoriť neskôr.

Začneme tým, že objasníme, prečo modelujeme náhodný výber tak, ako hovorí definícia 6.2.6 (vyplynie to z odpovede na bod (b)). Predpoklad nezávislosti veličín X_i je zaiste na mieste vtedy, keď výber prvkov tej série 100 výrobkov sa realizuje náhodným výberom s *vrátením*. To znamená, že po náhodnom vybratí prvého výrobku (a zistení či je, resp. nie je nepodarok) sa ten vybratý vráti do základného súboru a potom sa náhodne ťahá ďalší. Aj keď, pochopiteľne, výsledkom nášho pokusu je konkrétna 100-tica núl a jednotiek, napr.

$$0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1, \dots$$

chápeme tieto hodnoty ako *realizácie* náhodných veličín X_i (veď keby sme zopakovali náhodný výber, získali by sme inú 100-ticu núl a jednotiek). Pripomeňme, že X_i je indikátor nepodarku, teda X_i nadobúda hodnotu 1 práve vtedy, keď i-tý výrobok je nepodarok (inak X_i nadobúda hodnotu 0).

Mechanizmus náhodného výberu s *vrátením* zaručuje dve skutočnosti:

1. Veličiny $X_1, X_2, \dots, X_{100}$ sú nezávislé náhodné veličiny.
2. Všetky X_i majú rovnaké rozdelenie (vrátením vytiahnutého výrobku sa zrejme pomery v základnom súbore zakaždým obnovia). Teraz tým rozdelením je rozdelenie $\mathrm{A}(p)$ s tým istým (neznámym) parametrom p.

Tieto dve kľúčové vlastnosti, sú obsahom pojmu *náhodný výber* (definícia 6.2.6). Podotýkame, že niekedy nemôžeme realizovať výber s vrátením, pretože meranie na jednotke výberu je deštruktívnej povahy (ako napr. pri meraní životnosti výrobkov, alebo pevnosti nosníkov). V takých prípadoch je síce porušená nezávislosť X_i, avšak obyčajne rozsah náhodného výberu je podstatne menší ako rozsah základného súboru, a preto nepatrné porušenie nezávislosti môžeme zanedbať (pozri úlohu 6.6.1). Termín *náhodný výber* sa používa aj tam, kde sa v skutočnosti nič nevyberá, ale za tých istých podmienok sa n-krát (a nezávisle na sebe) meria sledovaná náhodná veličina X (pozri príklad 6.2.8).

6.2.6 Definícia. Náhodný výber z rozdelenia, ktoré určuje distribučná funkcia F, je n-tica náhodných veličín $X_1, X_2, \dots, X_n$ s vlastnosťami:

1. Všetky X_i majú rovnaké rozdelenie. Funkcia F je distribučnou funkciou každej X_i.
2. Veličiny $X_1, X_2, \dots, X_n$ sú nezávislé náhodné veličiny.

Číslo n sa nazýva *rozsah* náhodného výberu. *Realizáciou náhodného výberu* rozumieme n-ticu reálnych čísel $x_1, x_2, \dots, x_n$, kde x_i chápeme ako realizáciu veličiny X_i, $i = 1, 2, \dots, n$.

6.2.7 Poznámka. Prečo definícia začína slovami náhodný výber z *rozdelenia*, keď tento článok začínal tým, že náhodne vyberáme zo základného súboru? To preto, lebo cieľ štatistického skúmania je urobiť (čo najrozumnejší) úsudok o rozdelení, ktoré má skúmaná náhodná veličina (ten znak základného súboru, ktorý vyšetrujeme). Vieme, že pravdepodobnostné rozdelenie určuje jeho distribučná funkcia F, resp. jeho hustota f, resp. jeho pravdepodobnostná funkcia f. Preto zámerom je – na základe náhodného výberu – získať čo najlepšiu predstavu o funkcii F, resp. f.

Niekedy však máme skromnejší cieľ, pretože niekedy nám stačí odhad strednej hodnoty skúmaného rozdelenia, resp. okrem odhadu strednej hodnoty, aj odhad napr. smerodajnej odchýlky, prípadne odhad niektorého kvantilu. To záleží od kontextu.

Na druhej strane, odpovede na takéto otázky závisia od toho, čo o neznámom rozdelení vieme. Rozdelenie je síce neznáme, ale niekedy predsa len máme o ňom nejakú parciálnu informáciu. Napr. vieme, že ide o exponenciálne rozdelenie, ale parameter je neznámy (6.1.6), alebo ide o rozdelenie A(p), kde nepoznáme hodnotu p (6.2.5). Preto odteraz je naozaj dôležité pozorne si všímať, za akých predpokladov o (neznámom) rozdelení robíme úsudky o tých, či oných charakteristikách toho rozdelenia.

6.2.8 Príklad. Predpokladajme, že máme odhadnúť plošný obsah P štvorca so stranou dĺžky a. Máme k dispozícii hodnoty $x_1, x_2, \dots, x_{10}$, ktoré predstavujú výsledky 10-násobného nezávislého merania dĺžky a. Ako spracovať tieto dáta? Naporúdzi sú aspoň dva postupy:

Postup 1. Ak x_i aproximuje neznáme a, tak x_i^2 aproximuje a^2, a preto za odhad $\mathrm{P} = a^2$ vezmeme priemer z hodnôt $x_1^2, \dots, x_{10}^2$, teda hodnotu

$$\frac{1}{10}\sum_{i=1}^{10} x_i^2.$$

Postup 2. Z hodnôt $x_1, x_2, \dots, x_{10}$ vypočítame priemer a ten považujeme za odhad neznámej hodnoty a. Hodnotu $\mathrm{P} = a^2$ odhadneme druhou mocninou priemeru, t.j. hodnotou

$$\left(\frac{1}{10}\sum_{i=1}^{10} x_i\right)^2.$$

Keby sme odmietli štatistický model situácie, nedokážeme rozhodnúť, ktorý postup dáva presnejší odhad, pretože skutočnú hodnotu a^2 nepoznáme. Štatistický model znamená, že hore uvedené postupy chápeme ako uplatnenie dvoch náhodných veličín – hovoríme im *štatistiky* – ktoré sa opierajú o pojem náhodného výberu:

$$\text{prvou je } g(X_1, X_2, \dots, X_{10}) = \frac{1}{10}\sum_{i=1}^{10} X_i^2 \quad \text{a druhou je } h(X_1, X_2, \dots, X_{10}) = \left(\frac{1}{10}\sum_{i=1}^{10} X_i\right)^2,$$

kde $X_1, X_2, \dots, X_{10}$ modelujú meranie dĺžky a, $X_i = a + E_i$, pre $i = 1, 2, \dots, 10$. Náhodná veličina E_i modeluje chybu i-tého merania. Merania sa realizovali tak, aby sme 10-ticu $E_1, E_2, \dots, E_{10}$ mohli považovať za 10-ticu nezávislých veličín. Najčastejšie sa predpokladá, že E_i majú rozdelenie N(0, σ^2), ale niekedy menej – iba to, že $\mathrm{E}(E_i) = 0$ a $var(E_i)$ existuje.

Pojem náhodného výberu umožňuje urobiť analýzu vlastností uvedených štatistík, čo urobíme v ďalšej kapitole. V nasledujúcom článku začneme s tými najjednoduchšími odhadovacími štatistikami: výberovým priemerom a výberovým rozptylom.

6.3 Výberový priemer a výberový rozptyl

6.3.1 Základnou výberovou charakteristikou je *výberový priemer*. Ak $X_1, X_2, \dots, X_n$ je náhodný výber z rozdelenia, ktoré je určené funkciou F, tak výberový priemer je náhodná veličina

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Ak rozdelenie (určené funkciou F), má strednú hodnotu m a varianciu σ^2, tak zrejme pre všetky X_i, platí: $\mathrm{E}(X_i) = m$, $var(X_i) = \sigma^2$ (lebo rozdelenie všetkých X_i určuje funkcia F).

6.3.2 Veta. Ak $X_1, X_2, \dots, X_n$ je náhodný výber z rozdelenia, ktorého stredná hodnota sa rovná m a variancia sa rovná σ^2, tak pre výberový priemer $\bar{X}$ platí

$$\mathrm{E}(\bar{X}) = m \quad \text{a} \quad var(\bar{X}) = \frac{\sigma^2}{n}.$$

Dôkaz vety sa opiera o 4.2.3 a 4.2.5. Veta odzrkadľuje intuitívne chápanie základnej vlastnosti štatistiky $\bar{X}$, že priemer akceptujeme ako prirodzený odhad strednej hodnoty. Zároveň vidíme, že variancia (rozptyl) výberového priemeru je tým menšia, čím je rozsah výberu väčší. Uvedomme si, že pre smerodajnú odchýlku $\sigma(\bar{X})$, t.j. smerodajnú odchýlku výberového priemeru platí:

$$\sigma(\bar{X}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}, \text{ kde } \sigma \text{ je smerodajná odchýlka rozdelenia, z ktorého výber pochádza.}$$

6.3.3 Príklad. Predpokladajme, že náhodný výber rozsahu 9, je výberom z rozdelenia, ktorého stredná hodnota $m = 50$ a variancia $\sigma^2 = 16$. Akú strednú hodnotu a akú smerodajnú odchýlku má výberový priemer? Aký rozsah by musel mať náhodný výber, aby výberový priemer mal smerodajnú odchýlku $\sigma(\bar{X}) \le 1$?

Riešenie. Odpoveď na prvú otázku dáva veta 6.3.2, podľa ktorej ($m = 50$, $\sigma = 4$, $n = 9$)

$$\mathrm{E}(\bar{X}) = m = 50 \quad \text{a} \quad var(\bar{X}) = \frac{\sigma^2}{n} = \frac{16}{9}, \quad \text{t.j. } \sigma(\bar{X}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{9}} = \frac{4}{3}.$$

Ak má platiť $\sigma(\bar{X}) = \sqrt{\frac{\sigma^2}{n}} = \frac{4}{\sqrt{n}} \le 1$, zrejme $\sqrt{n} \ge 4$, odkiaľ pre hľadané n máme $n \ge 16$.

Zopakujme, že strednú hodnotu základného súboru odhadujeme výberovým priemerom. Ako odhadujeme varianciu, teda rozptyl základného súboru? Pravdepodobne každý pozná vzorec, ktorému hovoríme *výberová variancia* (= *výberový rozptyl*).

6.3.4 Nech $X_1, X_2, \dots, X_n$ je náhodný výber z rozdelenia, ktoré má strednú hodnotu m a varianciu σ^2. Za tohoto predpokladu definujeme výberovú varianciu S^2 vzťahom

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

Všimnime si, že výraz $\sum_{i=1}^{n}(X_i - \bar{X})^2$ môžeme upraviť na tvar $\sum_{i=1}^{n} X_i^2 - n\bar{X}^2$, s ktorým sa niekedy lepšie pracuje (pozri napr. dôkaz vety 7.1.11). Zrejme

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) = \sum_{i=1}^{n} X_i^2 - 2\bar{X} \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \bar{X}^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2,$$

pretože $\sum_{i=1}^{n} X_i = n\bar{X}$.

6.3.5 Poznámka. Dá sa ukázať, že výberová variancia je náhodná veličina, pre ktorú platí

$$E(S^2) = E\left(\frac{1}{n-1} \sum_{i=1}^{n}(X_i - \bar{X})^2 \right) = \sigma^2,$$

čo zdôvodňuje, prečo ňou odhadujeme varianciu rozdelenia (z ktorého výber pochádza). Podotýkame, že robíme tak vtedy, keď nič viac o rozdelení nevieme (iba to, že rozptyl má). Keby sme napr. vedeli, že ide o výber z exponenciálneho rozdelenia, postupovali by sme inak – o tom budeme hovoriť v nasledujúcej kapitole.

Ak máme odhadnúť smerodajnú odchýlku rozdelenia, použijeme štatistiku

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n}(X_i - \bar{X})^2}\,.$$

Realizáciu štatistiky S budeme označovať s. O tom, aké vlastnosti majú štatistiky S^2, resp. S (ako odhady variancie, resp. smerodajnej odchýlky), budeme hovoriť v kapitole 7.

6.3.6 Príklad. Predpokladajme, že náhodný výber rozsahu 11 je výberom z rozdelenia, ktoré má (neznámu) strednú hodnotu m a (neznámu) varianciu σ^2 (a nič viac o rozdelení nevieme). Realizáciu náhodného výberu predstavujú dáta

3.30 2.80 4.24 3.47 3.65 3.19 2.61 2.82 3.17 3.84 4.02

Určíme odhady strednej hodnoty m a smerodajnej odchýlky σ uvažovaného rozdelenia.

Riešenie. Vedecké kalkulačky majú základné štatistické funkcie, a tak stačí zistiť, ktorým tlačítkom získame realizáciu výberového priemeru, resp. výberovej smerodajnej odchýlky. Kalkulačky majú implementovaný výpočet s (často uvedený pod symbolom σ_{n-1}) a tiež verziu σ_n, keď v menovateli figuruje n a nie $n-1$.

V našom prípade sa výberový priemer realizuje hodnotou 3.3736, ktorú považujeme za odhad strednej hodnoty m. Výberová smerodajná odchýlka S sa realizuje hodnotou $s =$ 0.5264 (čo je odhad neznámej smerodajnej odchýlky). Pre úplnosť dodávame, že tlačítko σ_n dáva hodnotu 0.5019 (všetky výsledky sú zaokrúhlené na 4 desatinné miesta).

6.4 Reprezentácia dát: tabuľky početností, diagramy, histogramy

6.4.1 V tomto článku predpokladáme, že skúmaný znak základného súboru je kvantitatívnej povahy. Zopakujme, že znak chápeme (modelujeme) náhodnou veličinou X a tá, ako vieme, má nejaké rozdelenie. Získať o ňom čo najlepšiu predstavu – to je náš cieľ. Musíme to dokázať na základe náhodného výberu. Rozdelenie určuje jeho distribučná funkcia F, alebo tiež pravdepodobnostná funkcia f (ak X je diskrétna), resp. hustota f (ak X je spojitá). V tomto článku budeme hovoriť o tom, ako získame základnú predstavu o tvare funkcie f. V článku 6.5 budeme hovoriť o odhadoch distribučnej funkcie F.

6.4.2 Ak X je diskrétna veličina a máme k dispozícii realizáciu náhodného výberu primerane veľkého rozsahu, predstavu o tvare pravdepodobnostnej funkcie f získame vizualizovaním tabuľky relatívnych početností. O význame tabuľky relatívnych početností sme hovorili už v kapitole 2 (pozri príklad 2.1.7). Pripomeňme si kontext toho príkladu.

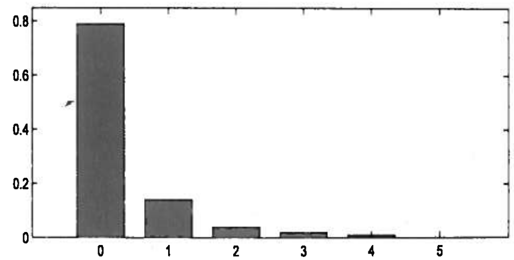
Automat zvára blatník auta na piatich miestach. Kontrolou 100 blatníkov sme zistili, že na 79 sa nenašiel ani jeden nekvalitný zvar, jeden nekvalitný zvar bol zistený na 14 blatníkoch, dva nekvalitné zvary na štyroch blatníkoch, tri na dvoch a štyri nekvalitné zvary iba v jednom prípade z tých 100 prípadov. Preto tabuľka relatívnych početností hodnôt veličiny X mala takúto podobu:

$H(X)$	0	1	2	3	4	5
Odhady $P(X = i)$	0.79	0.14	0.04	0.02	0.01	0

Realizácia náhodného výberu je 100-tica hodnôt x_i, kde x_i znamená počet nekvalitných zvarov na i-tom blatníku. Celú 100-ticu neuvádzame, ale povedzme, že začínala takto:

$$1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, \dots$$

O tejto 100-tici vieme, že absolútna početnosť hodnoty 0 sa rovná 79, resp. početnosť hodnoty 1 sa rovná 14, kým početnosť hodnoty 5 sa rovná 0. Situáciu chápeme tak, že tých 100 blatníkov bolo vyrobených za rovnakých podmienok a nezávisle jeden od druhého. X_1 predstavuje počet nekvalitných zvarov na prvom blatníku, t. j. X_1 je veličina X v prvom opakovaní „pokusu" a napr. X_{100} je veličina X v poslednom pokuse. Pretože ide o zložený pokus (100-násobné opakovanie pokusu za rovnakých podmienok) musíme sa na X_1 až X_{100} pozerať ako na 100 nezávislých kópií veličiny X (opäť opakujeme obsah pojmu náhodný výber). Pretože všetky X_1, X_2 až X_{100} majú rovnaké rozdelenie (dané funkciou f) realizácia výberu $X_1, X_2, \dots, X_{100}$ dáva informáciu o tvare funkcie f. Na obrázku 6-1 je stĺpcový diagram, ktorý vizualizuje tabuľku relatívnych početností.



Obr. 6-1. Stĺpcový diagram rozloženia relatívnych početností

Ak treba odhadnúť napr. $P(X \le 1)$, urobíme to relatívnou početnosťou, preto kladieme

$$P(X \le 1) \approx \frac{79+14}{100} = 0.93$$

6.4.3 Teraz hovorme o prípade, keď rozdelenie veličiny X určuje hustota f. Ak $X_1, X_2, \ldots, X_n$ je náhodný výber z rozdelenia, ktoré určuje funkcia f, tak analogicky ako v 6.4.2, realizácia náhodného výberu $x_1, x_2, \ldots, x_n$ poskytuje relatívnu početnosť udalosti $\{a \le X < b\}$. Konkrétne, povedzme, že napr. 23 údajov spomedzi $x_1, x_2, \ldots, x_{100}$ leží v intervale $\langle 10, 20)$. Ak je to tak, tak neznámu pravdepodobnosť $P(10 \le X < 20)$ aproximujeme relatívnou početnosťou udalosti $\{10 \le X < 20\}$

$$P(10 \le X < 20) \approx \frac{23}{100} = 0.23$$

Ak dáta usporiadame, tak hodnotu absolútnej početnosti stanovíme ľahko. V takýchto situáciách, samozrejme, využívame nejaké výpočtové prostredie. Pri tejto príležitosti poznamenávame, že v štatistike sa udomácnilo označenie $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, pre *usporiadaný* súbor, čo znamená, že hodnoty $x_1, x_2, \ldots, x_n$ sa usporiadajú do neklesajúcej postupnosti

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n-1)} \le x_{(n)}$$

a tak, napr. $x_{(1)} = \min(x_1, x_2, \ldots, x_n)$ a $x_{(n)} = \max(x_1, x_2, \ldots, x_n)$. Túto myšlienku získavania relatívnych početností na základe náhodného výberu urobíme koncepčnejšie tak, že

a) obor hodnôt veličiny X rozložíme na k disjunktných intervalov $I_1, I_2, \ldots, I_k$ (tried)
b) nájdeme absolútne, resp. relatívne početnosti jednotlivých intervalov (tried)
c) získané výsledky prehľadne tabelujeme
d) zostrojíme histogram, ktorý (zatiaľ len voľne povedané), zobrazuje tabuľku absolútnych, resp. relatívnych početností.

6.4.4 Príklad. Sledovaným znakom, teda meranou veličinou, je životnosť elektronickej súčiastky istého typu. Za rovnakých podmienok sa sledovala životnosť 150 takýchto súčiastok. Výsledkom merania je realizácia náhodného výberu rozsahu 150:

2.82	7.65	4.64	6.58	9.63	3.66	5.48	2.82	7.23	2.21
0.68	4.30	5.77	5.39	4.60	4.47	3.19	6.39	10.88	5.01

......

Všetkých 150 údajov nemá zmysel uvádzať. Pre lepšiu čitateľnosť dát, uveďme súbor už ako *usporiadaný* súbor, pričom hodnoty (v 100-kách hodín) sú zapísané postupne do stĺpcov, zhora dole a zľava doprava (napr. $x_{(4)} = 1.22$, $x_{(15)} = 1.86$, $x_{(16)} = 1.93$ atď.):

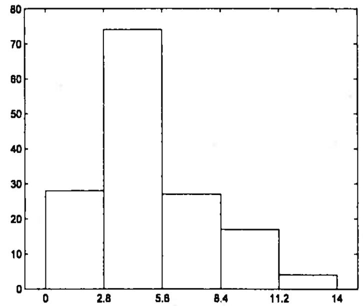
0.56	1.93	2.87	3.42	4.02	4.51	5.13	6.07	6.71	9.09
0.68	2.03	2.88	3.44	4.02	4.54	5.15	6.08	6.72	9.37
0.98	2.11	2.89	3.46	4.16	4.60	5.16	6.09	6.93	9.43
1.22	2.19	2.98	3.47	4.23	4.64	5.37	6.16	6.99	9.56
1.24	2.21	3.09	3.49	4.24	4.71	5.39	6.17	7.23	9.63
1.30	2.45	3.09	3.50	4.29	4.78	5.40	6.19	7.65	9.75
1.31	2.54	3.17	3.51	4.29	4.84	5.42	6.25	7.65	9.85
1.32	2.55	3.19	3.65	4.30	4.85	5.43	6.27	7.85	10.10
1.33	2.59	3.19	3.66	4.32	4.91	5.43	6.29	8.10	10.11

1.51	2.61	3.19	3.66	4.34	4.92	5.46	6.39	8.43	10.88
1.54	2.63	3.19	3.71	4.36	4.96	5.48	6.42	8.47	11.14
1.72	2.65	3.20	3.84	4.41	4.97	5.51	6.56	8.60	11.35
1.74	2.71	3.22	3.85	4.43	4.98	5.77	6.58	8.90	11.92
1.80	2.82	3.25	3.92	4.44	5.01	5.87	6.69	8.93	12.57
1.86	2.82	3.30	4.00	4.47	5.05	5.98	6.70	8.94	13.53

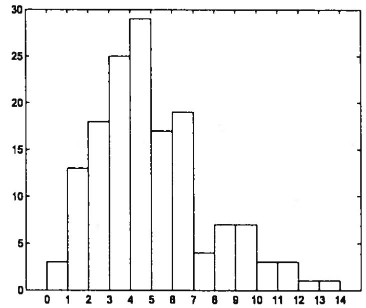
Ak zvolíme 5 tried, tak šírka triedy sa rovná 2.8 a početnosti tried uvádza tabuľka

hranice tried	0 – 2.8	2.8 – 5.6	5.6 – 8.4	8.4 – 11.2	11.2 – 14.0
početnosti	28	74	27	17	4

Zobrazenie rozloženia absolútnych početností dáva histogram na obr. 6-2. Zrejme 5 tried je málo, obrázok má malú výpovednú hodnotu, triedy sú príliš široké, a tak ich absolútne početnosti príliš zjednodušujú rozloženie tých 150 hodnôt.



Obr. 6-2. Počet tried sa rovná 5



Obr. 6-3. Počet tried sa rovná 14

Ak zvolíme 14 tried, triedy majú šírku 1 a teraz (pre zmenu) tabuľka namiesto hraníc tried, uvádza stredy tried, tzv. *triednych reprezentantov* x_j^*. Neskôr ukážeme ako ich využijeme pri aproximovaní výberových štatistík. Tu je tabuľka početností:

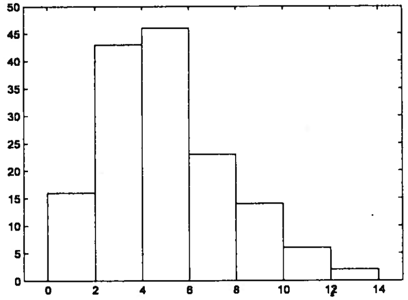
x_j^*	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5	10.5	11.5	12.5	13.5
n_j	3	13	18	25	29	17	19	4	7	7	3	3	1	1

Vizualizáciu tabuľky dáva obr. 6-3. Teraz je však tried príliš veľa a ako vidíme, tvar histogramu je preto trochu strapatý. Zdá sa, že optimálny počet tried môže byť napr. sedem. Vtedy šírka tried sa rovná 2 (a to je celkom pohodlné):

triedy:	0 – 2	2 – 4	4 – 6	6 – 8	8 – 10	10 – 12	12 – 14
početnosti:	16	43	46	23	14	6	2

V tomto prípade histogram dáva dobrú predstavu o tvare hustoty skúmanej veličiny, pozri obr. 6-4. Jednoznačne sa nedá povedať, pri koľkých triedach získavame najlepší výsledok, avšak existujú isté odporúčania. Asi historicky prvým je Sturgesovo pravidlo, keď odporúčaný počet tried k získame podľa vzťahu $k \approx 1 + 3.3\log_{10}(n)$.

6.4.5 Poznámka. Najčastejšie sa triedy volia rovnako široké a v takom prípade výšky obdĺžnikov reprezentujú absolútne, resp. relatívne početnosti tried. Je celkom jedno, či ide o absolútne, alebo o relatívne početnosti, tvar histogramu je v oboch prípadoch rovnaký. Niekedy sme však v situácii, keď originálne dáta už nie sú k dispozícii, a to, čo dostaneme ako výsledok experimentu, je tabuľka početností tried, pričom triedy nemajú rovnakú šírku. Vtedy je namieste otázka, kedy je diagram početností skutočným histogramom.



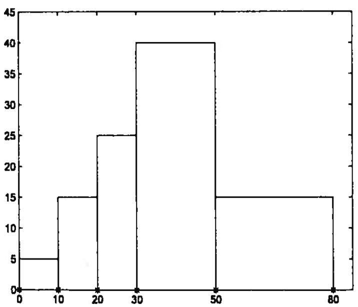
Obr. 6-4. Počet tried sa rovná 7 (príklad 6.4.4)

Za neformálnu definíciu správneho histogramu vezmime tento popis: Histogram je taký stĺpcový diagram, v ktorom *plochy* obdĺžnikov sú úmerné absolútnym (alebo relatívnym) početnostiam tried. Keď sú triedy rovnako široké, tak je to to isté, ako povedať, že výšky obdĺžnikov sú úmerné početnostiam tried, avšak pozor, ak triedy nie sú rovnako široké, tak nejde o to isté. Nasledujúci príklad ilustruje rozdiel.

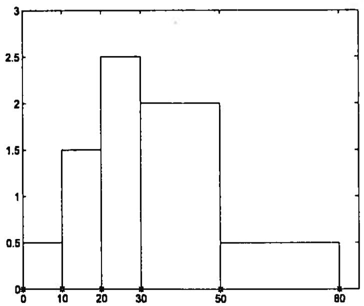
6.4.6 Príklad. Nasledujúca tabuľka predstavuje triedy a ich početnosti. Zostrojme správny histogram.

triedy	0 – 10	10 – 20	20 – 30	30 – 50	50 – 80
početnosti	5	15	25	40	15

Obrázok 6-5 predstavuje stĺpcový diagram, v ktorom výšky stĺpcov reprezentujú absolútne početnosti tried. To však nie je histogram. Histogram je na obrázku 6-6, lebo tu absolútne početnosti tried sú úmerné plochám obdĺžnikov. Vidíme, že tvary diagramov sú rôzne!



Obr. 6-5. Stĺpcový diagram



Obr. 6-6. Histogram

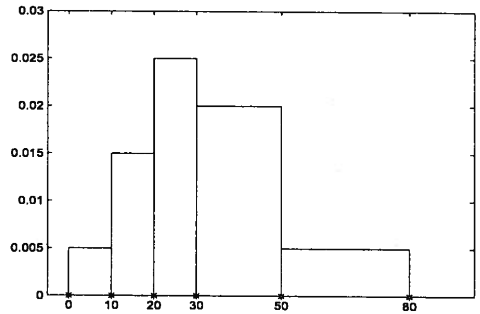
Ako sme získali výšky stĺpcov jednotlivých obdĺžnikov na obrázku 6-6? Podľa poznámky 6.4.5 pre výšky v_i má platiť:

$$10\,v_1 = k\,5, \qquad 10\,v_2 = k\,15, \qquad 10\,v_3 = k\,25 \qquad 20\,v_4 = k\,40, \qquad 30\,v_5 = k\,15,$$

kde k je koeficient úmernosti (ľubovoľné kladné číslo). Vzali sme najjednoduchšiu voľbu, $k = 1$ a získali sme $v_1 = 0.5$, $v_2 = 1.5$, $v_3 = 2.5$, $v_4 = 2$, $v_5 = 0.5$. Niekedy si prajeme zostrojiť histogram tak, aby súčet plôch všetkých obdĺžnikov sa rovnal jednej. Napríklad preto, lebo ho chceme porovnať s grafom (do úvahy prichádzajúcej) hustoty. V takom prípade majú platiť hore uvedené vzťahy, kde k už nie je ľubovoľné, pretože má platiť:

$$10\,v_1 + 10\,v_2 + 10\,v_3 + 20\,v_4 + 30\,v_5 = k\,5 + k\,15 + k\,25 + k\,40 + k\,15 = k\,100 = 1, \text{ t. j. } k = 0.01,$$

a preto máme $v_1 = 0.005$, $v_2 = 0.015$, $v_3 = 0.025$, $v_4 = 0.020$, $v_5 = 0.005$.



Obr. 6-7. Histogram, v ktorom súčet plôch obdĺžnikov sa rovná 1

6.5 Výberová (empirická) distribučná funkcia a výberové kvantily

6.5.1 V tomto článku pôjde o konštrukciu *výberovej* distribučnej funkcie F_n. Namiesto výberová sa niekedy hovorí *empirická*, pretože realizáciu $x_1, x_2, \dots, x_n$ náhodného výberu $X_1, X_2, \dots, X_n$ chápeme ako empirické údaje. Výber pochádza z rozdelenia, ktoré určuje (neznáma) distribučná funkcia F. Naším cieľom je z dát $x_1, x_2, \dots, x_n$ zostrojiť F_n tak, aby sme hodnoty $F_n(x)$ mohli považovať za dobré odhady neznámych $F(x)$.

Predstavíme dve verzie empirickej distribučnej funkcie. Prvú označíme symbolom G_n a bude to funkcia nespojitá, po častiach konštantná, definovaná celkom prirodzeným spôsobom. Druhá verzia, ktorú budeme označovať ako F_n, bude spojitá a po častiach lineárna a budeme ňou určovať výberové kvantily.

V tomto odseku preberieme konštrukciu funkcie G_n. Už v 6.4.2, resp. v 6.4.3, sme na základe náhodného výberu neznámu pravdepodobnosť $P(a \le X < b)$ odhadovali relatívnou početnosťou udalosti $\{\, a \le X < b \,\}$. Keďže teoretická distribučná funkcia F je definovaná vzťahom $F(x) = P(X < x)$, za odhad hodnoty $F(x)$ je prirodzené zobrať relatívnu početnosť udalosti $\{X < x\}$. To znamená, že hodnotu $G_n(x)$ definujeme vzťahom

$$G_n(x) = \frac{card\{\, x_i : x_i < x \,\}}{n}.$$

Takto je funkcia G_n skutočne výberovým protipólom (teoretickej) distribučnej funkcie F.

Ak namiesto súboru dát $x_1, x_2, \ldots, x_n$ pracujeme s usporiadaným súborom $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, tak uvedený definitorický vzťah môžeme prepísať takto:

$G_n(x) = 0$, pre $x \le x_{(1)}$, lebo pre $x \le x_{(1)}$, zrejme $card\{x_i : x_i < x\} = card(\emptyset) = 0$,

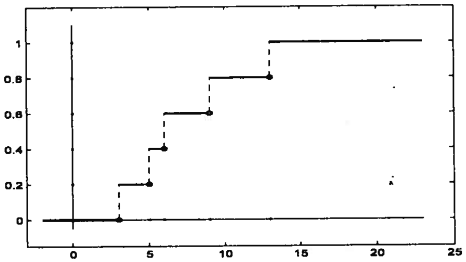
$G_n(x) = \frac{i}{n}$, pre $x_{(i)} < x \le x_{(i+1)}$, pretože pre tieto x zrejme $card\{x_i : x_i < x\} = i$,

$G_n(x) = 1$, pre $x > x_{(n)}$, keďže pre takéto x, $card\{x_i : x_i < x\} = n$.

Zrejme

$$P(a \le X < b) \approx \frac{card\{x_i : a \le x_i < b\}}{n} = \frac{card\{x_i : x_i < b\} - card\{x_i : x_i < a\}}{n} = G_n(b) - G_n(a).$$

Obrázok 6-8 dáva graf G_n pre (usporiadaný) dátový vektor [3, 5, 6, 9, 13].



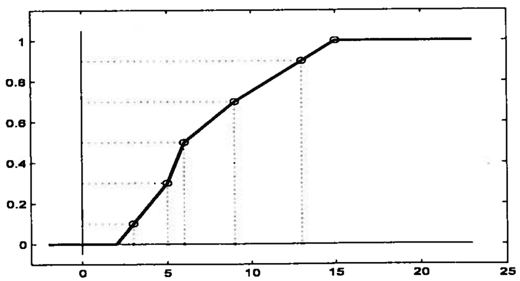
Obr. 6-8. Graf po častiach konštantnej empirickej distribučnej funkcie G_n

6.5.2 Druhú verziu empirickej (t. j. výberovej) distribučnej funkcie budeme označovať F_n. Jej výhodou je, že nám pohodlne poskytuje výberové kvantily, resp. výberové kvartily, resp. akýkoľvek výberový p-kvantil. Funkcia F_n je spojitá a po častiach lineárna. Budeme ju definovať tak, že popíšeme konštrukciu jej grafu.

Najprv spojme úsečkami dvojice bodov $[x_{(i)}, \frac{2i-1}{2n}]$, $[x_{(i+1)}, \frac{2i+1}{2n}]$ pre $i = 1, \ldots, n-1$. Nech

$$a = x_{(1)} - \frac{x_{(2)} - x_{(1)}}{2}, \text{ resp. } b = x_{(n)} + \frac{x_{(n)} - x_{(n-1)}}{2}.$$

Spojme úsečkou bod $[a, 0]$ s bodom $[x_{(1)}, \frac{1}{2n}]$ a bod $[x_{(n)}, \frac{2n-1}{2n}]$ s bodom $[b, 1]$. Nakoniec, nech $F_n(x) = 0$, ak $x \le a$, resp. $F_n(x) = 1$, ak $x \ge b$. Takýmto spôsobom graf funkcie F_n zobrazuje interval $\langle a, b\rangle$ (ktorý obsahuje všetky dáta x_i) na interval $\langle 0, 1\rangle$ (na osi y). Obrázok 6-9 dáva graf F_n pre (usporiadaný) dátový vektor [3, 5, 6, 9, 13].



Obr. 6-9. Graf spojitej, po častiach lineárnej, empirickej distribučnej funkcie F_n

Na osi x (obr. 6-9) sú vyznačené body usporiadaného súboru: $x_{(1)}, \ldots, x_{(5)}$, t. j. 3, 5, 6, 9, 13. Na osi y uvažujeme osnovu bodov: $\frac{2i-1}{2n}$, pre $i = 1, 2, \ldots, 5$, t.j. body $\frac{1}{10}, \frac{3}{10}, \frac{5}{10}, \frac{7}{10}, \frac{9}{10}$. Takto dostaneme 5 bodov grafu: [3, 0.1], [5, 0.3], [6, 0.5], [9, 0.7], [13, 0.9]. Body spojíme úsečkami. Určíme a, b: $a = 3 - 1 = 2$, $b = 13 + 2 = 15$ a podľa návodu dokončíme graf F_n.

6.5.3 Pripomeňme, že p-kvantil rozdelenia, ktoré určuje distribučná funkcia F, je taký bod x_p, v ktorom má funkcia F hodnotu p, teda pre teoretický p-kvantil x_p platí: $F(x_p) = p$. Preto analogicky, *výberovým kvantilom* bude taký bod $\tilde{x}_p$, že

$$F_n(\tilde{x}_p) = p.$$

Z obrázku 6-9 vieme vyčítať, že napr.

výberovým 0.5-kvantilom je číslo 6, t. j. $\tilde{x}_{0.5} = 6$, pretože $F_n(6) = 0.5$,

výberovým 0.8-kvantilom je číslo 11, t. j. $\tilde{x}_{0.8} = 11$, lebo $F_n(11) = 0.8$,

výberovým 0.9-kvantilom je číslo 13, t. j. $\tilde{x}_{0.9} = 13$, lebo $F_n(13) = 0.9$.

Na určenie *výberového mediánu* $\tilde{x}_{0.5}$, nie je nutné zostrojiť F_n, stačí pracovať s usporiadaným výberovým súborom. Ak počet prvkov n výberového súboru je nepárny, tak výberový medián $\tilde{x}_{0.5}$ definujeme ako prostredný údaj, resp. ak rozsah výberu n je párne číslo, tak $\tilde{x}$ definujeme ako priemer dvoch „prostredných" dát. Formálnejšie

$$\tilde{x}_{0.5} = \begin{cases} x_{(k+1)}, & ak\ n = 2k+1, \\ \dfrac{x_{(k)} + x_{(k+1)}}{2}, & ak\ n = 2k. \end{cases}$$

Analogicky ako v kapitole 4 pri teoretickom medzikvartilovom rozpätí, výberové medzikvartilové rozpätie mkr definujeme ako vzdialenosť medzi výberovým horným kvartilom $\tilde{x}_{0.75}$ (často sa označuje q_3) a výberovým dolným kvartilom $\tilde{x}_{0.25}$ ($= q_1$)

$$mkr = \tilde{x}_{0.75} - \tilde{x}_{0.25} = q_3 - q_1.$$

6.5.4 Príklad. Nech náhodná veličina X má rozdelenie, ktoré určuje distribučná funkcia F. Predpokladajme, že nám nie je nič známe o tvare funkcie F a na základe náhodného výberu rozsahu 13, ktorý sa realizoval hodnotami:

3.4 1.1 1.9 1.8 4.6 3.9 2.7 1.3 2.2 2.4 1.6 1.5 3.0

máme odhadnúť $P(2 < X < 3)$, určiť výberový medián a výberové medzikvartilové rozpätie.

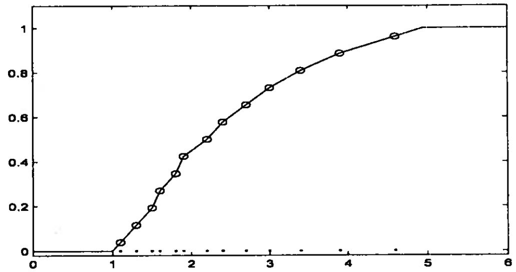
Riešenie. Samozrejme, v reálnej situácii, v ktorej nevieme nič o tvare rozdelenia, potrebujeme nie 13, ale oveľa viac dát. Teraz nám ide len o ilustrovanie základných výberových pojmov.

Najprv realizáciu výberu usporiadajme (v reálnych situáciách je nutné použiť nejaké výpočtové prostredie, pretože rozsah výberov je aj niekoľkonásobne väčší):

1.1 1.3 1.5 1.6 1.8 1.9 2.2 2.4 2.7 3.0 3.4 3.9 4.6

$P(2 < X < 3)$ odhadneme relatívnou početnosťou, t. j. číslom $\frac{3}{13}$, pretože 3 dáta ležia v (2, 3). Výberový medián sa rovná 2.2, pretože 6 údajov leží naľavo a 6 napravo od hodnoty 2.2.

Na určenie výberových kvantilov pre $p = 0.25$ a $p = 0.75$, je dobré zostrojiť F_n.



Osnovu na osi x vytvárajú dané dáta: 1.1, 1.3, 1.5, 1.6, ... Osnovu na osi y tvoria body $\frac{1}{26}, \frac{3}{26}, \frac{5}{26}, \frac{7}{26}, \ldots$, ktoré (v desatinnom rozvoji a po zaokrúhlení) sa rovnajú postupne 0.0385, 0.1154, 0.1923, 0.2692, ... Pretože 0.25 (ide o dolný kvartil!) leží medzi 0.1923 a 0.2692, tak 0.25-kvantil $\tilde{x}_{0.25}$ leží medzi $x_{(3)}$ a $x_{(4)}$, t. j. medzi 1.5 a 1.6. Presnú hodnotu nájdeme lineárnym interpolovaním, resp. jednoducho uplatnením úmernosti:

$$\tilde{x}_{0.25} = 1.5 + x, \text{ kde pre } x \text{ platí: } \frac{x}{0.25 - 0.1923} = \frac{1.6 - 1.5}{0.2692 - 0.1923}. \text{ Takto } \tilde{x}_{0.25} = 1.5 + 0.075 = 1.575.$$

Výberový 0.75-kvantil $\tilde{x}_{0.75}$ stanovíme analogicky. Bodu $x_{(10)}$ odpovedá na osi y bod $\frac{2\cdot 10-1}{26} = \frac{19}{26} \approx 0.7308$ a bodu $x_{(11)}$ odpovedá bod $\frac{2\cdot 11-1}{26} = \frac{21}{26} \approx 0.8077$. Pretože 0.75 patrí intervalu (0.7308, 0.8077), $\tilde{x}_{0.75}$ zrejme leží medzi $x_{(10)} = 3.0$ a $x_{(11)} = 3.4$. Preto $\tilde{x}_{0.75}$ určíme ako

$$\tilde{x}_{0.75} = 3.0 + x, \text{ kde pre } x \text{ platí: } \frac{x}{0.75 - 0.7308} = \frac{3.4 - 3.0}{0.8077 - 0.7308}. \text{ Takto } \tilde{x}_{0.75} = 3.0 + 0.0999 \approx 3.100.$$

Pre výberové medzikvartilové rozpätie máme

$$mkr = 3.100 - 1.575 = 1.525.$$

6.5.5 Príklad. Merala sa dĺžka bezporuchovej práce elektronického zariadenia (v hodinách). Tabuľka dáva výsledky 150 meraní v tvare tried a ich absolútnych početností.

triedy	0 – 300	300 – 600	600 – 900	900 – 1200	1200 – 1700	1700 – 2500
početnosti	10	18	30	36	36	20

a) Aproximujme hodnotu výberového priemeru a výberovej smerodajnej odchýlky.
b) Aproximujme hodnotu výberového mediánu a výberového medzikvartilového rozpätia.

Riešenie. Súčet tých (neznámych) hodnôt, ktoré patria prvej triede môžeme aproximovať hodnotou 1500, pretože 150 je reprezentant (t. j. stred) prvej triedy a 10 je jej početnosť. Rovnako súčet tých (neznámych) hodnôt, ktoré patria druhej triede môžeme aproximovať hodnotou $18 \cdot 450 = 8100$, pretože 450 je reprezentant (stred) druhej triedy a 18 je početnosť tej triedy. Označme stredy tried (reprezentantov) x_j^* a početnosti tried n_j, $j = 1, \ldots, k$. Potom máme

$$\sum_{i=1}^n x_i \approx n_1 x_1^* + n_2 x_2^* + \cdots + n_k x_k^* = \sum_{j=1}^k n_j x_j^*, \text{ a preto } \bar{x} \approx \frac{1}{n} \sum_{j=1}^k n_j x_j^*.$$

Takto pre dané dáta máme

$$\bar{x} = \frac{1}{150}[10 \cdot 150 + 18 \cdot 450 + 30 \cdot 750 + 36 \cdot 1050 + 36 \cdot 1450 + 20 \cdot 2100] = \frac{164100}{150} = 1094.$$

Analogicky

$$\sum_{i=1}^n (x_i - \bar{x})^2 \approx \sum_{j=1}^k n_j (x_j^* - \bar{x})^2, \text{ a preto } s^2 \approx \frac{1}{n-1} \sum_{j=1}^k n_j (x_j^* - \bar{x})^2.$$

Pre naše dáta

$$s^2 = \frac{1}{149}[10 \cdot 891136 + 18 \cdot 414736 + \cdots + 20 \cdot 1012036] = \frac{44799600}{149} = 300668.4564,$$

odkiaľ $s \approx 548.33$.

Výberový medián $\tilde{x}$ stanovíme lineárnou interpoláciou. Najprv overme, že $\tilde{x}$ leží v triede (900, 1200). Kumulatívna početnosť hodnôt do hranice 900 sa rovná 58 (= 10 + 18 + 30), a preto relatívna početnosť sa rovná 58/150 = 0.387 (zaokrúhľujeme na tri desatinné miesta). Kumulatívna relatívna početnosť hodnôt do hranice 1200 sa rovná 94/150 = 0.627. Pretože hodnota $p = 0.5$ (ide o výberový medián) leží v intervale (0.387, 0.627), medián je skutočne bodom intervalu (900, 1200). Jeho presnú hodnotu získame lineárnym interpolovaním:

$$\tilde{x} = 900 + x, \text{ kde pre } x \text{ platí: } \frac{x}{0.50 - 0.387} = \frac{1200 - 900}{0.627 - 0.387}. \text{ Takto } \tilde{x} = 900 + 141.25 \approx 1041.$$

Aproximácie kvartilov získame úplne analogicky. Dolný kvartil identifikujeme v triede (600, 900) a pre aproximáciu $\tilde{x}_{0.25}$ dostávame (po zaokrúhlení) hodnotu 695. Horný kvartil je bodom triedy (1200, 1700) a aproximácia $\tilde{x}_{0.75}$ sa (po zaokrúhlení) rovná 1457. Preto pre výberové medzikvartilové rozpätie máme

$$mkr = 1457 - 695 = 762.$$

6.6 Úlohy

6.6.1 Uvažujme o malej modifikácii situácie z príkladu 6.1.3. V osudí máme tentoraz iba 10 lístkov (s číslami 1 až 10). Uvažujme o náhodnom výbere rozsahu 5, ktorý modelujeme 5-ticou $X_1, X_2, \ldots, X_5$, čo znamená, že X_1 predstavuje číslo na prvom vytiahnutom lístku, X_2 je číslo na druhom vytiahnutom, ... atď. Rozlišujme dva prípady:

a) Najprv nech ide o náhodný výber s vrátením. Ukážte, že v takomto prípade
   - všetky X_i majú to isté rozdelenie, $X_i \sim R\{1, 2, \ldots, 10\}$,
   - veličiny $X_1, X_2, \ldots, X_5$ sú nezávislé.

b) Teraz nech ide o výber bez vrátenia. Ukážte, že v tomto prípade
   - opäť všetky X_i majú to isté rozdelenie, $X_i \sim R\{1, 2, \ldots, 10\}$,
   - ale veličiny $X_1, X_2, \ldots, X_5$ sú závislé.

Nakoniec si uvedomme, že keby v osudí nebolo 10, ale bolo napr. 200 lístkov, tak závislosť v prípade výberu rozsahu 5 *bez vrátenia* je taká slabá, že ju môžeme zanedbať. Túto skutočnosť pri modelovaní reálnych situácií naozaj využívame, keď z nejakých dôvodov nie je možné realizovať výber s vrátením (viď odsek pred definíciou 6.2.6).

6.6.2 Pokračujme v diskusii o situácii „Tombola", ale teraz ako o štatistickej úlohe (príklad 6.1.4). Cieľom je (na základe výsledku náhodného výberu) odhadnúť počet lístkov v osudí. Priznávame, že jednoduchšie bude pracovať za predpokladu, že ide o *výber s vrátením*, keď modelom výberu sú *nezávislé* veličiny X_1, X_2, X_3, X_4, X_5. Vieme, že majú rovnomerné rozdelenie na množine $\{1, 2, ..., N\}$, kde N je neznáme prirodzené číslo (počet lístkov v osudí). Ide teda o situáciu, keď síce poznáme *typ* rozdelenia, ktoré majú pozorované veličiny X_i, ale nepoznáme parameter toho rozdelenia. Parameter N máme odhadnúť na základe realizácie $x_1, x_2, ..., x_5$ náhodného výberu.

Uvažujme takto: 5-krát (a nezávisle) pozorujeme X, pričom $X \sim R\{1, 2, \ldots, N\}$. Hoci jej strednú hodnotu vieme určiť, ostáva neznámou, pretože závisí od neznámeho N. K dispozícii máme náhodný výber, resp. jeho realizáciu a vieme (podľa 6.3.2), že *výberový priemer* je prirodzený odhad strednej hodnoty. Položte do rovnosti nájdenú strednú hodnotu $E(X)$ na jednej strane a výberový priemer (ako štatistiku) na druhej strane a zo získanej rovnosti vyjadrite odhad neznámeho parametra N. Akú konkrétnu hodnotu (pre odhad N) dáva získaný odhad pri dátach 11, 7, 15, 8, 4 (z príkladu 6.1.4)?

6.6.3 Zatrieďte 150 dát z príkladu 6.4.4 do piatich tried: 0 – 2, 2 – 4, 4 – 7, 7 – 10, 10 – 14 (zľava uzavretých, sprava otvorených), čo nebude ťažké, pretože je k dispozícii už usporiadaný súbor. Uveďte výsledok v tvare tabuľky triednych početností. Znázornite odpovedajúci histogram.

6.6.4 Uvažujte opäť o realizácii náhodného výberu z príkladu 6.4.4 a vezmite ako fakt, že realizácia výberového priemeru má hodnotu 4.9871 a realizácia výberovej smerodajnej odchýlky sa rovná 2.6455.

a) Uvažujte o zatriedení dát do siedmych tried šírky 2: 0 – 2, 2 – 4, 4 – 6, ... , 12 – 14, ktoré je urobené v rámci príkladu 6.4.4 a vypočítajte jednak aproximáciu výberového priemeru a tiež výberovej smerodajnej odchýlky z tabuľky triednych početností, akoby ste originálne dáta nemali k dispozícii.

b) Teraz vezmite zatriedenie do 14 tried šírky 1 (aj to je urobené v rámci 6.4.4) a opäť vyčíslite aproximácie uvažovaných výberových charakteristík.

Očakávate, že v prípade 14 tried dostanete lepšie aproximácie ako v prípade siedmych? Asi áno, ale zrejme nie s istotou. Ako to dopadlo v tomto konkrétnom prípade?

6.6.5 Dáta predstavujú realizáciu náhodného výberu rozsahu $n = 11$ z príkladu 6.2.3

118, 112, 110, 113, 115, 109, 100, 133, 129, 105, 121

a) Vypočítajte realizácie výberového priemeru a výberovej smerodajnej odchýlky.
b) Nájdite hodnoty skokovej, po častiach konštantnej výberovej distribučnej funkcie $G_n(x)$, pre $x = 110, 111, 121, 125$.
c) Nájdite hodnoty spojitej, po častiach lineárnej výberovej distribučnej funkcie $F_n(x)$, pre $x = 110,\ 111,\ 121, 125$.
d) Určte hodnotu výberového mediánu a výberového medzikvartilového rozpätia.

7 BODOVÉ ODHADY

7.1 Výberové rozdelenia, nevychýlenosť odhadu

7.1.1 V tejto kapitole – na rozdiel od predchádzajúcej kapitoly – formulujeme východisko našich úvah takto:

> Skúmaná náhodná veličina X, resp. skúmané rozdelenie (ktoré má veličina X) má známu distribučnú funkciu F, ale predpis funkcie F obsahuje parameter θ, ktorého hodnota nie je známa. Na základe náhodného výberu, resp. na základe jeho realizácie, máme nájsť čo najlepší odhad parametra θ.

Laicky vzaté, neznámu hodnotu θ chceme dobre, resp. čo najpresnejšie odhadnúť. To znamená, že chceme nájsť postup, ktorým keď spracujeme naše dáta (realizáciu náhodného výberu), získame číslo, ktoré bude ležať blízko neznámeho bodu θ. Uvedomme si však, že tieto vety sú veľmi neurčité, veď čo znamená *najpresnejšie* odhadnúť, čo znamená *ležať blízko* bodu θ, keď θ predstavuje neznámy bod nejakého intervalu? Týmto formuláciám treba dať vecnejší obsah. Začneme upozornením, že termín odhad bude mať dva významy. Prvým významom bude konkrétna hodnota. Za odhad neznámeho bodu θ vezmeme akési konkrétne číslo. Dostaneme ho tak, že nejakým prirodzeným, resp. intuitívne obhájiteľným spôsobom spracujeme dáta – realizáciu náhodného výberu. Druhým významom termínu odhad bude náhodná veličina, prostredníctvom ktorej sme spracovali dáta a vypočítali konkrétnu hodnotu. Teda pod druhým významom termínu odhad rozumieme náhodnú veličinu, ktorá je funkciou náhodného výberu $X_1, X_2, \ldots, X_n$. Túto veličinu nazývame *výberová štatistika*, resp. *odhadovacia štatistika* a budeme ju označovať symbolom $h(X_1, X_2, \ldots, X_n)$. Cieľom kapitoly je objasniť, čo znamená dobre odhadnúť neznámy parameter θ.

7.1.2 Príklad. Uvažujme rozdelenie $R(0, \theta)$, v ktorom θ je neznámy parameter. Nájdime nejakú, resp. nejaké odhadovacie štatistiky na odhad θ.

Riešenie. Prirodzený odhad vzniká takto: Stredná hodnota $E(X)$ uvažovaného rozdelenia sa rovná $\frac{\theta}{2}$ (článok 4.1). Z článku 6.3 zase vieme, že odhadom strednej hodnoty je štatistika výberový priemer, a tak ak $\bar{X}$ odhaduje $\frac{\theta}{2}$, potom $2\bar{X}$ odhaduje θ, to znamená, že kladieme

$$\hat{\theta} = 2\bar{X} = \frac{2}{n}\sum_{i=1}^{n} X_i\,, \quad \text{teda} \quad h(X_1, X_2, \ldots, X_n) = 2\bar{X}.$$

Hovoríme, že uplatňujeme *momentovú* metódu, pretože teoretický moment dávame do rovnosti s výberovým momentom (v tomto prípade strednú hodnotu sme dali do rovnosti s výberovým priemerom). Iný odhad poskytuje štatistika $max(X_1, X_2, \ldots, X_n)$, lebo je prirodzené považovať maximálne x_i, teda $x_{(n)}$, za „temer dobrý" odhad θ. Žiada sa, pravda, trochu zväčšiť hodnotu $x_{(n)}$, aby sme sa dostali bližšie k θ (o takej modifikácii budeme hovoriť neskôr). Pre konkrétnosť, predpokladajme, že máme k dispozícii náhodný výber rozsahu 10:

0.5, 1.4, 0.3, 1.8, 2.4, 0.7, 1.1, 0.1, 1.6, 1.9

Odhad $2\bar{X}$ sa realizuje hodnotou 2·1.18 = 2.36. Tomuto odhadu nedôverujeme, pretože najväčší údaj sa rovná 2.4 a je zrejmé, že θ nemôže byť menšie ako niektoré x_i. Druhý odhad

$$g(X_1, X_2, \dots, X_{10}) = max(X_1, X_2, \dots, X_{10})$$

(zatiaľ neupravovaný), dáva hodnotu 2.4. Tento konkrétny odhad môžeme považovať (zatiaľ laicky a len v tejto situácii) za lepší. Samozrejme, cítime, že je nutné formulovať vlastnosti, ktoré má mať dobrý odhad (keď naň nazeráme ako na odhadovaciu štatistiku).

7.1.3 Príklad. Predpokladajme, že konkrétnou technológiou vyrábané elektronické komponenty majú životnosť, ktorú modelujeme exponenciálnym rozdelením Exp(λ), v ktorom parameter λ *nie je známy.* Z celkovej produkcie náhodne vyberieme 20 výrobkov a na základe merania ich životností, t. j. na základe nameraných hodnôt $x_1 = 855$, $x_2 = 1048$, ..., $x_{20} = 1792$, chceme odhadnúť

a) strednú hodnotu životnosti,
b) pravdepodobnosť toho, že životnosť (náhodne vybratého) výrobku prekročí 1200 hodín.

Riešenie. Nech veličina X predstavuje *životnosť* vyrábaných výrobkov. Podľa predpokladu $X \sim \text{Exp}(\lambda)$ a v bode (a) chceme odhadnúť E(X).

Z článku 6.3 vieme, že odhadom strednej hodnoty E(X) je výberový priemer $\bar{X}$. Ak napr. z daných 20 dát sa výberový priemer realizuje hodnotou 1243, tak za odhad E(X) vezmeme 1243. Situácia je trochu iná ako v príklade 7.1.2, kde išlo priamo o odhad parametra. Teraz ide o odhad E(X) a z kapitoly 4 vieme, že $E(X) = \frac{1}{\lambda}$. To znamená, že odhadovať E(X) znamená, odhadovať hodnotu *známej* funkcie *neznámeho* λ.

Ešte viditeľnejšie je to v otázke (b), kde ide o odhad $P(X > 1200)$. Z kapitoly 2 vieme, že pri exponenciálnom rozdelení pravdepodobnosť $P(X > 1200) = e^{-\lambda\, 1200}$. To znamená, že aj pri odhade $P(X > 1200)$ ide o odhad (známej) funkcie neznámeho parametra λ.

Stojíme pred úlohou odhadnúť hodnoty *známych* funkcií neznámeho λ. Akceptujme intuitívny postup – momentovou metódou získame odhad $\hat{\lambda}$ neznámeho parametra λ (pozri 7.1.2) a ak treba odhadnúť hodnotu známej funkcie τ neznámeho parametra λ, teda treba odhadnúť $\tau(\lambda)$, za odhad vezmime $\tau(\hat{\lambda})$.

Realizujme tento postup na získanie odhadu $P(X > 1200)$. Momentovou metódou nájdeme odhad λ tak, že z rovnosti $\frac{1}{\lambda} = \bar{X}$ vyjadríme λ (na ľavej strana rovnosti je teoretická stredná hodnota E(X), na pravej strane – jej výberový protipól.) Takto sme dostali štatistiku na odhad λ – je ňou

$$\hat{\lambda} = h(X_1, X_2, \dots, X_n) = \frac{1}{\bar{X}}.$$

Pre konkrétnosť, ak sa výberový priemer z 20 dát realizoval hodnotou 1243, odhadom λ je číslo $\frac{1}{1243} \approx 0.0008$ (po zaokrúhlení).

Náš cieľ je nájsť odhad pre $P(X > 1200)$. Keďže $P(X > 1200) = e^{-\lambda\, 1200}$, za odhad neznámej hodnoty $e^{-\lambda\, 1200}$ vezmeme štatistiku $e^{-\hat{\lambda}\, 1200}$ (za λ sme dosadili jej odhad). Konkrétne,

$$\text{odhad } P(X > 1200) = e^{-\hat{\lambda}\, 1200} = e^{-0.0008\, 1200} = e^{-0.96} = 0.383.$$

Formálnejšie, za odhad $P(X > 1200)$ sme vzali štatistiku $g(X_1, X_2, \dots, X_n) = e^{-1200/\bar{X}}$.

7.1.4 Poznámka. V príkladoch 7.1.2 a 7.1.3 sme odhady, resp. odhadovacie štatistiky získali prirodzeným spôsobom. Je na mieste opýtať sa, aké dobré sú tieto odhady, resp. kedy vôbec odhadovaciu štatistiku považujeme za dobrú. Alebo, na základe čoho rozhodneme, ktorá z tých dvoch štatistík v príklade 7.1.2 je lepším odhadom.

Zrejme je potrebné formulovať vlastnosti, ktoré žiadame od „dobrých" odhadov, t. j. od dobrých odhadovacích štatistík. Základnou vlastnosťou odhadu je (ne)vychýlenosť. Pripomeňme, že pod frázou *rozdelenie určuje funkcia* $f(.\,;\theta)$ rozumieme to, že rozdelenie určuje pravdepodobnostná funkcia f (ak X je diskrétna), resp. rozdelenie určuje hustota f (ak X je spojitá s hustotou f). V oboch prípadoch θ predstavuje neznámy parameter funkcie f.

7.1.5 Definícia. Nech $X_1, X_2, \dots, X_n$ je náhodný výber z rozdelenia, ktoré je určené funkciou $f(.\,;\theta)$. Štatistiku $h(X_1, X_2, \dots, X_n)$ nazývame *nevychýlený odhad* parametra θ, ak pre všetky do úvahy prichádzajúce θ platí

$$E[\, h(X_1, X_2, \dots, X_n)\,] = \theta.$$

Ak štatistika $g(X_1, X_2, \dots, X_n)$ má odhadovať hodnoty funkcie $\tau(\theta)$ (τ je známa funkcia, θ je neznámy parameter), tak $g(X_1, X_2, \dots, X_n)$ je nevychýlený odhad $\tau(\theta)$, ak pre všetky do úvahy prichádzajúce θ platí

$$E[\, g(X_1, X_2, \dots, X_n)\,] = \tau(\theta).$$

Voľne povedané, nevychýlenosť odhadu znamená, že odhad nemá systematickú chybu. Keby sme napr. sledovali hodnoty nevychýleného odhadu v dlhom rade (nezávislých) experimentov, tak ťažisko jeho realizácií by sa zhodovalo s bodom θ, resp. s bodom $\tau(\theta)$.

7.1.6 Príklad. Ukážme, že v príklade 7.1.2, v ktorom išlo o odhady parametra θ v rozdelení R(0, θ), je odhad $h(X_1, X_2, \dots, X_n) = 2\bar{X}$ nevychýleným odhadom θ.

Riešenie. Ak $X_1, X_2, \dots, X_n$ je náhodný výber z R(0, θ), tak $E(X_i) = \theta/2$, pre $i = 1, \dots, n$. Preto

$$E[\, 2\bar{X}\,] = 2\,E[\, \tfrac{1}{n} \textstyle\sum_{i=1}^{n} X_i\,] = \tfrac{2}{n} \sum_{i=1}^{n} E(X_i) = \tfrac{2}{n}\, n\, \tfrac{\theta}{2} = \theta.$$

To znamená, že v dlhom rade prípadov, odhad $2\bar{X}$ nemá systematickú chybu, avšak, videli sme (pozri 7.1.2), že pri konkrétnej realizácii tento odhad dal hodnotu, ktorej sme neuverili (veď $x_5 = 2.4$, tak odhadu $\hat{\theta} = 2.36$ nedôverujeme). Vidíme, že vlastnosť nevychýlenosti ešte nezaručuje, že odhad je naozaj dobrý. Odhad je totiž náhodnou veličinou, a tak okrem jej strednej hodnoty je potrebné všímať si aj jej ďalšie vlastnosti, napr. varianciu. Tá totiž spôsobila, že konkrétna realizácia odhadu sa rovnala 2.36, teda hodnote, ktorú nechceme akceptovať.

Urobme (za pomoci výpočtového prostredia) simulačný experiment. Zvoľme za θ hodnotu 2.5 a generujme 1000 výberov rozsahu 10 z rozdelenia R(0, 2.5). V každom výbere vyčíslime uvažovanú štatistiku $2\bar{X}$. Tých 1000 realizácií štatistiky $2\bar{X}$ zobrazme histogramom a získame tak predstavu o jej rozdelení. V MATLABe to realizujú príkazy:

```
>>  x = 2.5*rand(10, 1000);
>>  y = 2*mean(x);
>>  hist(y, 16)
```

Histogram na obr. 7-1 nám dáva predstavu o rozdelení štatistiky $2\bar{X}$.

7.1.7 Príklad. Uvažujme o odhade $max(X_1, X_2, \ldots, X_n)$ z príkladu 7.1.2. Nie je ťažké ukázať, že pre $E[\,max(X_1, X_2, \ldots, X_n)\,]$ platí (pomôcka: príklad 3.3.11)

$$E[\,max(X_1, X_2, \ldots, X_n)\,] = \frac{n}{n+1}\,\theta.$$

Tento výsledok znamená, že $max(X_1, X_2, \ldots, X_n)$ je vychýlený odhad θ. Modifikujme ho tak, aby sme získali nevychýlený odhad θ.

Riešenie. Ak $E[\,max(X_1, X_2, \ldots, X_n)\,] = \frac{n}{n+1}\theta$, tak $\frac{n+1}{n}E[\,max(X_1, X_2, \ldots, X_n)\,] = \theta$. Operátor strednej hodnoty je lineárny, a preto

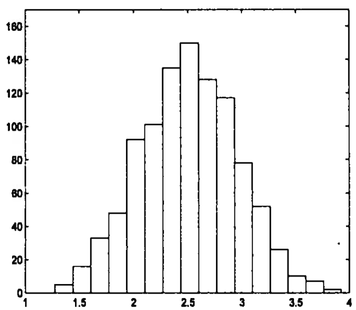
zo vzťahu $\frac{n+1}{n}E[\,max(X_1, X_2, \ldots, X_n)\,] = \theta$, máme $E[\,\frac{n+1}{n}max(X_1, X_2, \ldots, X_n)\,] = \theta$.

To znamená, že štatistika

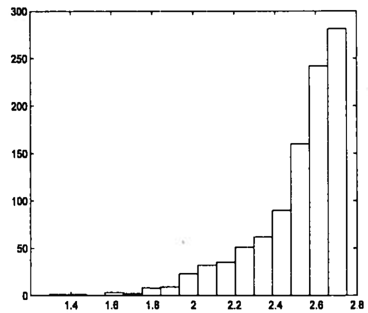
$$g(X_1, X_2, \ldots, X_n) = \frac{n+1}{n}max(X_1, X_2, \ldots, X_n)$$

je nevychýlený odhad parametra θ. Pre zaujímavosť, realizácia tejto štatistiky (z 10-tich dát z príkladu 7.1.2) sa rovná (11/10)·2.4 = 2.64. Na základe tohoto jediného číselného výsledku nemôžeme posúdiť kvalitu získaného odhadu, pretože skutočnú hodnotu parametra nepoznáme. Môžeme však získať predstavu o rozdelení štatistiky $g(X_1, X_2, \ldots, X_n)$ rovnako, ako v 7.1.6. Generujme 1000 výberov rozsahu 10 z rozdelenia R(0, 2.5) a 1000-krát vyčíslime uvažovanú štatistiku. V MATLABe to dosiahneme príkazmi

```
>>  x = 2.5*rand(10, 1000);
>>  y = 11*max(x)/10;
>>  hist(y, 16)
```



Obr. 7-1. Histogram rozdelenia štatistiky $2\bar{X}$



Obr. 7-2. Histogram štatistiky $\frac{11}{10}max(X_1, \ldots, X_{10})$

7.1.8 Poznámka. Sumarizujme obsah príkladov 7.1.2, 7.1.6 a 7.1.7. Ide o odhad parametra θ v rozdelení R(0, θ). Diskutujeme o dvoch odhadoch:

$$h(X_1, X_2, \ldots, X_n) = \frac{2}{n}\sum_{i=1}^{n} X_i \quad \text{a} \quad g(X_1, X_2, \ldots, X_n) = \frac{n+1}{n}max(X_1, X_2, \ldots, X_n).$$

Vieme už, že obidva sú nevychýlené odhady θ. Je na mieste otázka: Ktorý z nich je lepší? Aj keď sme simuláciou získali predstavu o rozdelení týchto štatistík (takéto rozdelenia nazývame *výberové rozdelenia*), netrúfame si – len na základe histogramov – dať jednoznačnú odpoveď. K situácii sa vrátime v článku 7.2 a tam otázku zodpovieme.

7.1.9 Príklad. Vyšetrime nevychýlenosť odhadov z príkladu 6.2.8, v ktorom ide o odhad plošného obsahu P štvorca, na základe meraní dĺžky strany a (toho štvorca). Uvažujeme o dvoch štatistikách

prvou je $g(X_1, X_2, \ldots, X_{10}) = \frac{1}{10}\sum_{i=1}^{10} X_i^2$ a druhou je $h(X_1, X_2, \ldots, X_{10}) = \left(\frac{1}{10}\sum_{i=1}^{10} X_i\right)^2$.

Ktorá z nich je nevychýleným odhadom plošného obsahu P?

Riešenie. Pripomeňme, že veličiny X_i modelujú (predstavujú) merania dĺžky a, t. j. pre $i = 1, 2, \ldots, 10$, máme: $X_i = a + E_i$, kde náhodná veličina E_i modeluje chybu i-tého merania. Merania sme realizovali tak, že $E_1, E_2, \ldots, E_{10}$ môžeme považovať za 10-ticu nezávislých veličín, pričom predpokladáme, že $E(E_i) = 0$ a $var(E_i) = \sigma^2$. Zrejme pre $i = 1, 2, \ldots, 10$ platí

$$E(X_i) = E(a + E_i) = a + E(E_i) = a + 0 = a,$$
$$E(X_i^2) = \sigma^2 + a^2,$$

pretože $var(X_i) = var(a + E_i) = var(E_i) = \sigma^2$, ale tiež $var(X_i) = E(X_i^2) - (EX_i)^2 = E(X_i^2) - a^2$. Konečne,

$$E[\,\tfrac{1}{10}\textstyle\sum_{i=1}^{10} X_i^2\,] = \tfrac{1}{10}\sum_{i=1}^{10} E(X_i^2) = \tfrac{1}{10}\sum_{i=1}^{10}(\sigma^2 + a^2) = \tfrac{1}{10}10(\sigma^2 + a^2) = \sigma^2 + a^2.$$

$$E[\,\left(\tfrac{1}{10}\textstyle\sum_{i=1}^{10} X_i\right)^2\,] = var(\tfrac{1}{10}\sum_{i=1}^{10} X_i) + (E[\,\tfrac{1}{10}\sum_{i=1}^{10} X_i\,])^2 = \tfrac{1}{100}10\sigma^2 + a^2 = \tfrac{\sigma^2}{10} + a^2.$$

Oba odhady sú vychýlené, ale vychýlenie odhadu $h(X_1, X_2, \ldots, X_{10})$ je menšie. Keď namiesto rozsahu 10, uvažujeme rozsah n, tak pre strednú hodnotu odhadu $h(X_1, X_2, \ldots, X_n)$ máme

$$E\left[\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i\right)^2\right] = var\left(\tfrac{1}{n}\sum_{i=1}^{n} X_i\right) + \left(E[\tfrac{1}{n}\sum_{i=1}^{n} X_i]\right)^2 = \tfrac{1}{n^2}n\sigma^2 + a^2 = \tfrac{\sigma^2}{n} + a^2.$$

Ak rozsah výberu n konverguje do nekonečna, tak vychýlenie odhadu $h(X_1, X_2, \ldots, X_n)$ konverguje k nule. Hovoríme, že odhad $h(X_1, X_2, \ldots, X_n)$ je *asymptoticky* nevychýlený.

7.1.10 V tomto článku diskutujeme základnú vlastnosť bodových odhadov – nevychýlenosť. Preto zopakujme to, čo sme povedali už v článku 6.3, na okraj odhadov strednej hodnoty a variancie (tentoraz aj s dôkazmi). Stredná hodnota je základná číselná charakteristika a ak ide o rozdelenie, v ktorom figuruje parameter θ, tak obyčajne $E(X)$ závisí na θ. Niekedy je $E(X)$ priamo rovné θ (napr. pri Poissonovom rozdelení), inokedy $E(X)$ je známou funkciou θ (napr. pri exponenciálnom rozdelení, $E(X) = 1/\theta$). Vo vete 7.1.11 nie je dôležité akou funkciou parametra θ je stredná hodnota $E(X)$, resp. akou funkciou parametra θ je variancia $var(X)$. Strednú hodnotu označme m, resp. varianciu označme σ^2 a môžeme ich považovať za nové parametre.

7.1.11 Veta. Ak $X_1, X_2, \ldots, X_n$ je náhodný výber z rozdelenia, ktorého stredná hodnota sa rovná m, tak výberový priemer $\bar{X}$ je nevychýlený odhad strednej hodnoty m, pretože

$$E(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i) = m.$$

Ak rozdelenie má varianciu σ^2, t. j. $var(X_i) = \sigma^2$, pre $i = 1, 2, \ldots, n$, tak výberová variancia S^2 je nevychýlený odhad variancie σ^2, pretože

$$E[\,\tfrac{1}{n-1}\textstyle\sum_{i=1}^{n}(X_i - \bar{X})^2\,] = \sigma^2.$$

Dôkaz. Prvá časť vety je jednoduchá. Využijeme vlastnosti strednej hodnoty z vety 4.2.3.

$$\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n} m = \frac{1}{n}\, n\, m = m.$$

V dôkaze druhej časti vety využijeme rovnosť $\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$, ktorú sme odvodili v 6.3.4. Počítajme

$$\mathrm{E}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \mathrm{E}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = \mathrm{E}\left(\sum_{i=1}^{n} X_i^2\right) - n\mathrm{E}(\bar{X}^2) = \sum_{i=1}^{n}\mathrm{E}(X_i^2) - n\mathrm{E}(\bar{X}^2).$$

Teraz potrebujeme určiť $\mathrm{E}(X_i^2)$ a $\mathrm{E}(\bar{X}^2)$. V oboch prípadoch použijeme známy vzťah

$$var(X) = \mathrm{E}(X^2) - (\mathrm{E}X)^2, \text{ v podobe } \mathrm{E}(X^2) = var(X) + (\mathrm{E}X)^2.$$

Najprv za X vezmeme X_i (na výpočet $\mathrm{E}(X_i^2)$) a potom $\bar{X}$, na výpočet $\mathrm{E}(\bar{X}^2)$. Takto

$$\mathrm{E}(X_i^2) = var(X_i) + m^2 = \sigma^2 + m^2, \quad \text{resp.} \quad \mathrm{E}(\bar{X}^2) = var(\bar{X}) + (\mathrm{E}\bar{X})^2 = \frac{\sigma^2}{n} + m^2.$$

Nakoniec,

$$\mathrm{E}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \sum_{i=1}^{n}\mathrm{E}(X_i^2) - n\mathrm{E}(\bar{X}^2) = n(\sigma^2 + m^2) - n\left(\frac{\sigma^2}{n} + m^2\right) = (n-1)\,\sigma^2,$$

odkiaľ už máme

$$\mathrm{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \sigma^2.$$

Zrejme štatistika $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ je síce vychýlený odhad σ^2, ale je aspoň asymptoticky nevychýlený, pretože

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \frac{1}{n}(n-1)\,\sigma^2 = \left(1 - \frac{1}{n}\right)\sigma^2.$$

7.2 Stredná kvadratická chyba odhadu

7.2.1 V ďalšom budeme štatistiku $h(X_1, X_2, \ldots, X_n)$ označovať symbolom $h(\boldsymbol{X})$. Ak $h(\boldsymbol{X})$ je odhadovacia štatistika parametra θ, tak je prirodzené nazývať náhodnú veličinu

$$h(\boldsymbol{X}) - \theta$$

chybou odhadu $h(\boldsymbol{X})$, resp. náhodnú veličinu

$$(h(\boldsymbol{X}) - \theta)^2$$

nazývať *kvadratickou chybou* odhadu $h(\boldsymbol{X})$. Je to stále náhodná veličina a vôbec nemusí byť jednoduché nájsť jej rozdelenie. Kvalitu odhadu však chceme charakterizovať len číslom, a preto vezmime *strednú hodnotu z kvadratickej chyby*: $\mathrm{E}[(h(\boldsymbol{X}) - \theta)^2]$. Odhad $h(\boldsymbol{X})$ je tým lepší, čím je táto stredná hodnota menšia. V štatistických učebniciach sa pre ňu udomácnilo označenie MSE(h) (Mean Square Error).

7.2.2 Definícia. Ak $h(\boldsymbol{X})$ je odhadovacia štatistika parametra θ, tak stredná kvadratická chyba odhadu h, MSE(h) je definovaná vzťahom

$$\mathrm{MSE}(h) = \mathrm{E}[(h(\boldsymbol{X}) - \theta)^2].$$

Ak $h(\boldsymbol{X})$ je nevychýlený odhad θ, tak $\mathrm{MSE}(h) = var(h(\boldsymbol{X}))$. Totiž nevychýlenosť znamená, že $\mathrm{E}[h(\boldsymbol{X})] = \theta$, a preto

$$\mathrm{MSE}(h) = \mathrm{E}[(h(\boldsymbol{X}) - \theta)^2] = \mathrm{E}[(h(\boldsymbol{X}) - \mathrm{E}[h(\boldsymbol{X})])^2] = var(h(\boldsymbol{X})).$$

Odhad $h(\boldsymbol{X})$ je tým lepší, čím je jeho MSE(h) menšia. Ak je $h(\boldsymbol{X})$ nevychýlený odhad θ, tak jeho stredná kvadratická chyba, MSE(h), sa rovná jeho variancii – $var(h(\boldsymbol{X}))$. To znamená, že o kvalite nevychýleného odhadu rozhoduje $var(h(\boldsymbol{X}))$. To je, samozrejme, v úplnej zhode s intuíciou.

Ak teda $h(\boldsymbol{X})$ a $g(\boldsymbol{X})$ sú dva nevychýlené odhady θ, tak za lepší z nich považujeme ten, ktorý má menšiu varianciu.

7.2.3 Príklad. V príkladoch 7.1.6, 7.1.7 sme ukázali, že tak $g(\boldsymbol{X}) = \frac{n+1}{n} max(X_1, X_2, \ldots, X_n)$ ako aj $h(\boldsymbol{X}) = 2\bar{X}$ sú nevychýlené odhady parametra θ v rozdelení R(0, θ). Ktorý z nich považujeme za lepší odhad θ?

Riešenie. Podľa 7.2.2 lepší z dvoch nevychýlených odhadov je ten, ktorý má menšiu varianciu. Použijeme fakt, že $var(max(X_1, X_2, \ldots, X_n)) = \frac{n\theta^2}{(n+1)^2(n+2)}$ (pomôcka: príklad 3.3.11).

Počítajme

$$var(g(\boldsymbol{X})) = var\left(\frac{n+1}{n} max(X_1, \ldots, X_n)\right) = \frac{(n+1)^2}{n^2} var(max(X_1, \ldots, X_n)) = \frac{(n+1)^2}{n^2}\,\frac{n\theta^2}{(n+1)^2(n+2)} = \frac{\theta^2}{n(n+2)}$$

$$var(h(\boldsymbol{X})) = var(2\bar{X}) = 4var(\bar{X}) = 4\,\frac{1}{n}\,\frac{\theta^2}{12} = \frac{\theta^2}{3n}$$

Keďže pre $n \geq 2$ zrejme $\frac{1}{n(n+2)} < \frac{1}{3n}$, variancia odhadu $g(\boldsymbol{X})$ je menšia ako variancia $h(\boldsymbol{X})$.

Preto odhad $\frac{n+1}{n} max(X_1, X_2, \ldots, X_n)$ považujeme za lepší ako odhad $2\bar{X}$.

7.2.4 Poznámka. Práve sme ukázali, že vieme rozhodnúť, ktorý z dvoch uvažovaných nevychýlených odhadov je lepší. Rovnako by sme vybrali najlepší spomedzi troch, či štyroch nevychýlených odhadov. Pokiaľ sa obmedzíme na nevychýlené odhady, tak nájsť najlepší nevychýlený odhad znamená nájsť ten z nich, ktorého variancia je (spomedzi všetkých nevychýlených odhadov) najmenšia možná. Táto problematika však už presahuje rámec tohoto textu. Nakoniec poznamenajme, že mierne vychýlený odhad $h(\boldsymbol{X})$ nie je treba ignorovať, pokiaľ jeho MSE(h) je menšia ako variancia konkurenčného nevychýleného odhadu, pretože MSE(h) je vskutku dobrý ukazovateľ kvality odhadu.

7.2.5 Príklad. Nech $X_1, X_2, \ldots, X_n$ je náhodný výber z rozdelenia N(μ, σ^2). Diskutujme o odhadoch parametrov μ a σ^2, vo svetle poznámky 7.2.4.

Riešenie. Začnime odhadom parametra μ. Vieme, že μ sa rovná strednej hodnote, a preto je celkom prirodzené odhadovať μ štatistikou $\bar{X}$ – *výberovým priemerom*. Veta 7.1.11 (tiež veta 6.3.2) hovoria, že výberový priemer je nevychýlený odhad E(X), teda nevychýlený odhad parametra μ. Je však na mieste otázka, či nie je možné odhadovať μ výberovým mediánom, keďže μ je nielen strednou hodnotou, ale aj mediánom rozdelenia N(μ, σ^2). Pretože diskutovanie vlastností štatistiky výberový medián je nad naše sily, vezmime bez dôkazu fakt, že výberový priemer je najlepší nevychýlený odhad parametra μ.

Teraz hovorme o odhadoch parametra σ^2, resp. σ. Pretože σ^2 sa rovná variancii, odhadovať σ^2, znamená odhadovať varianciu. Ak pritom poznáme μ, tak môžeme uvažovať o štatistike

$$S^2(\mu) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

ako o prirodzenom odhade variancie. Nie je ťažké ukázať, že je to nevychýlený odhad σ^2.

Dá sa ukázať viac (čo je však nad naše sily), totiž, že je to najlepší nevychýlený odhad parametra σ^2. Častejšia je však situácia, keď treba odhadovať σ^2, a pritom μ nepoznáme. Z vety 7.1.11 vieme, že *výberová variancia*

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

je nevychýlený odhad variancie, teda nevychýlený odhad σ^2. Dá sa ukázať, že je to najlepší nevychýlený odhad σ^2. Pre zaujímavosť poznamenávame, že štatistika

$$M_2 = \frac{1}{n} \sum_{i=1}^{n}(X_i - \bar{X})^2$$

ktorá – ako vieme – je síce vychýlený odhad σ^2, má menšiu strednú kvadratickú chybu ako S^2. To je dôvod, prečo niekedy dávame prednosť M_2 pred S^2.

Nakoniec hovorme o odhade smerodajnej odchýlky σ. Z poznámky 6.3.5 vieme, že na odhad σ použijeme štatistiku

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

Teraz dodávame, že je to *vychýlený* odhad parametra σ (dôkaz je obsahom úlohy 7.3.4). Dôležité je uvedomiť si, že celý odsek 7.2.5 sa vzťahuje na prípad, keď skúmané rozdelenie je normálne rozdelenie $N(\mu, \sigma^2)$.

7.3 Úlohy

7.3.1 Nech $X_1, X_2, \dots, X_n$ je náhodný výber z rozdelenia $N(\mu, \sigma^2)$, pričom neznámym parametrom je iba σ^2 (μ je teda známa hodnota). Ukážte, že štatistika

$$S^2(\mu) = \frac{1}{n} \sum_{i=1}^{n}(X_i - \mu)^2$$

je nevychýlený odhad σ^2.

7.3.2 Nech X_1, X_2, X_3, X_4 je náhodný výber z rozdelenia $R(0, \theta)$. Uvažujme o štatistikách

$$g(\boldsymbol{X}) = X_1 + \frac{1}{2}X_2 + \frac{1}{3}X_3 + \frac{1}{6}X_4, \qquad h(\boldsymbol{X}) = \frac{1}{2}(X_1 + X_2 + X_3 + X_4).$$

Ukážte, že obe štatistiky sú nevychýlené odhady parametra θ. Rozhodnite, ktorú z nich považujeme za lepší odhad θ (zdôvodnite).

7.3.3 Nech $X_1, X_2, \dots, X_{15}$ je náhodný výber z rozdelenia, ktorého stredná hodnota sa rovná m a variancia sa rovná σ^2. Uvažujme o štatistikách

$$h(\boldsymbol{X}) = \frac{1}{15}(X_1 + X_2 + \dots + X_{15}), \qquad g(\boldsymbol{X}) = \frac{1}{2}\left[\frac{1}{10}\sum_{i=1}^{10} X_i + \frac{1}{5}\sum_{i=11}^{15} X_i\right].$$

Ukážte, že obe štatistiky sú nevychýlené odhady strednej hodnoty m. Ktorej z nich dáme prednosť pri odhade m? Prečo?

7.3.4 Predpokladajme, že štatistika $h(\boldsymbol{X})$ je nevychýlený odhad parametra θ. Ukážte, že ak $var(h(\boldsymbol{X}))$ je kladná, tak štatistika $(h(\boldsymbol{X}))^2$ je vychýlený odhad θ^2. Uvážte, že z tohto tvrdenia vyplýva, že štatistika S je *vychýlený* odhad smerodajnej odchýlky σ.

7.3.5 Nech $X_1, X_2, \dots, X_n$ je náhodný výber z $R(-\theta, \theta)$. Stanovte k tak, aby štatistika

$$h(\boldsymbol{X}) = \frac{k}{n}(X_1^2 + X_2^2 + \dots + X_n^2)$$

bola nevychýlený odhad θ^2.

8 TESTOVANIE ŠTATISTICKÝCH HYPOTÉZ

Začnime objasnením, čo je štatistická hypotéza. Môže ňou byť hypotéza o rozdelení skúmanej náhodnej veličiny X, napríklad, hypotéza môže deklarovať, že rozdelenie veličiny X je normálne. Na základe náhodného výberu z rozdelenia veličiny X máme rozhodnúť, či hypotézu zamietneme, alebo nezamietneme. Inou a veľmi častou je situácia, v ktorej síce poznáme tvar rozdelenia skúmanej veličiny X, ale vystupuje v ňom neznámy parameter. Obsahom hypotézy je v takom prípade tvrdenie o hodnote parametra. Ďalším príkladom štatistickej hypotézy môže byť tvrdenie, že skúmané veličiny X a Y majú rovnakú strednú hodnotu, alebo, v celkom inej situácii, hypotézou môže byť tvrdenie, že skúmané veličiny X a Y sú nezávislé. Toto je len prvý (zatiaľ neúplný) náčrt úloh, s ktorými sa budeme zaoberať v tejto kapitole.

Možné formulácie testov budeme preberať postupne a základné pojmy objasňujeme najprv v rámci konkrétnych, jednoduchých situácií. Vzhľadom na vymedzený priestor, táto kapitola je len elementárny úvod do naozaj širokej štatistickej disciplíny, ktorá sa do dnešnej podoby sformovala v priebehu prvej polovice 20. storočia.

8.1 Test jednoduchej hypotézy proti jednoduchej alternatíve

8.1.1 Predpokladajme, že rozdelenie skúmanej veličiny X je dané funkciou f, ale predpis funkcie f obsahuje parameter θ, ktorého hodnota je neznáma. V tomto článku rozoberieme proces, v ktorom máme rozhodnúť medzi dvomi možnosťami: buď parameter θ sa rovná známej hodnote θ_0, alebo sa rovná inej (tiež známej) hodnote θ_1. Rozhodnutie máme urobiť na základe konkrétnej, v experimente získanej realizácie $x_1, x_2, \dots, x_n$ náhodného výberu $X_1, X_2, \dots, X_n$ (z rozdelenia veličiny X). V tejto kapitole budeme n-tice $x_1, x_2, \dots, x_n$ chápať ako body x v priestore R^n a písať $x = (x_1, x_2, \dots, x_n)$.

Na chvíľu ignorujme našu konkrétnu realizáciu $x = (x_1, x_2, \dots, x_n)$ náhodného výberu. Uvažujme o tom, aké (potenciálne možné) realizácie x sú pravdepodobné v prípade, že platí $\theta = \theta_0$ a naopak, aké sú pravdepodobné v prípade, keď v skutočnosti $\theta = \theta_1$. Keďže poznáme predpis pre funkciu f, vieme to (po chvíli počítania) zistiť. Tie realizácie x, ktoré sú veľmi *nepravdepodobné* v prípade, keď $\theta = \theta_0$, obrazne povedané, „vyzbierame" a práve nimi vytvoríme oblasť K. Oblasť K nazývame *kritická* oblasť, lebo vlastná realizácia testu pobeží podľa pravidla:

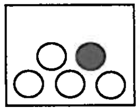
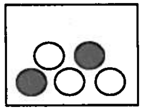
Možnosť $\theta = \theta_0$ zamietame práve vtedy, ak naša (v experimente získaná) realizácia $x \in$K.

Toto pravidlo treba akceptovať – je to dohodnutý postup, je to súčasť štatistického testu, lebo takto (a nie inak) bude prebiehať rozhodovanie. Musíme akceptovať aj isté označenia, a to, čo bolo voľne povedané v predchádzajúcich vetách, dostane formálnejšiu podobu. Možnosť $\theta = \theta_0$ budeme označovať ako *nulová hypotéza* a prípad $\theta = \theta_1$ označujeme ako *alternatívna hypotéza*. Test budeme zapisovať v tvare:

$$H_0 : \theta = \theta_0 \quad \text{proti} \quad H_1 : \theta = \theta_1.$$

V tomto článku sú obe hypotézy jednoduché, čo znamená, že obe deklarujú iba jednu hodnotu parametra. Priznávame, že v praxi zriedkakedy proti sebe stoja dve jednoduché hypotézy. Ale aj táto jednoduchá situácia dovoľuje objasniť základné pojmy a myšlienky testovania. Vykonať test znamená urobiť rozhodnutie – zamietnuť H_0, alebo nezamietnuť H_0. Už vieme, že H_0 zamietame práve vtedy, keď realizácia x náhodného výberu sa nachádza v kritickej oblasti. To znamená, že *test je definovaný kritickou oblasťou.* Dobre testovať, znamená prijať dobré rozhodnutie. Urobiť dobré rozhodnutie znamená optimálne stanoviť kritickú oblasť. Tieto trochu vágne formulácie je treba upresniť. Pomôžeme si nasledujúcimi príkladmi. (Pri druhom čítaní tohoto článku sa zamyslite nad tým, prečo píšeme o „dobrom" rozhodovaní a nemôžeme písať o *bezchybnom* rozhodovaní).

8.1.2 Príklad. V škatuli (do ktorej pochopiteľne nevidíme) máme 5 loptičiek. Povedzme, že do úvahy pripadajú len dve možnosti:

Buď škatuľa obsahuje: [4 biele a 1 čiernu] alebo [3 biele a 2 čierne].

Máme rozhodnúť, ktorá z týchto možností je správna, na základe náhodného výberu (výberu s vrátením) rozsahu 3 (tak malý rozsah volíme zatiaľ len pre jednoduchosť). Tri razy náhodne ťaháme loptičku, zistíme jej farbu a vrátime ju späť. Pokus má zrejme 8 možných výsledkov. Po vykonaní výberu máme rozhodnúť, či sa prikloníme k tvrdeniu, že škatuľa obsahuje jednu čiernu loptičku, alebo k tvrdeniu, že obsahuje dve čierne loptičky.

Formulujme túto lapidárnu situáciu ako štatistický test. Ak náhodne ťaháme loptičku (ktorá je buď biela, alebo čierna), tak máme do činenia s náhodnou veličinou X, ktorá má alternatívne rozdelenie. Dohodnime sa, že X bude indikátor čiernej, t. j. $P(X=1)=p$ (p je pravdepodobnosť vytiahnutia čiernej). V takom prípade, ak v škatuli je päť loptičiek a len jedna je čierna, tak $X \sim A(p)$, $p=\frac{1}{5}=0.2$, kým v druhom prípade (ak medzi piatimi sú dve čierne), $p=\frac{2}{5}=0.4$. Test je testom o parametri alternatívneho rozdelenia a má tvar

$$H_0: p=0.2 \quad \text{proti} \quad H_1: p=0.4$$

Ak X_i je indikátor čiernej v i-tom ťahu, tak náhodný výber rozsahu 3 predstavuje vektor (X_1, X_2, X_3) a súčet $S=X_1+X_2+X_3$ znamená počet čiernych medzi tromi vybratými. Zrejme $S \sim \mathrm{Bi}(3, p)$. Veličinu S nazývame *testovacia štatistika*, pretože nesie informáciu o parametri p. Prečo? Lebo malé hodnoty S korešpondujú s prípadom, keď $p=0.2$, t. j. platnou hypotézou je H_0, kým veľké hodnoty S hovoria v prospech H_1. Ako sme povedali, možným výsledkom náhodného výberu je niektorá z ôsmych trojíc:

$$(0,0,0)\ (1,0,0)\ (0,1,0)\ (0,0,1)\ (1,1,0)\ (1,0,1)\ (0,1,1)\ (1,1,1).$$

Výsledok $x=(1,1,1)$ môžeme považovať za *kritický* pre H_0, veď hovorí jasne v prospech H_1. Preto vezmime za kritickú oblasť K v R^3 polpriestor

$$K=\{(x_1,x_2,x_3)\in R^3 : x_1+x_2+x_3 \geq 3\},$$

ktorý z hore uvedených ôsmych výsledkov, obsahuje iba bod $x=(1,1,1)$. Stanovením kritickej oblasti sme zostrojili test, lebo deklarovaním K je definované naše rozhodovanie:

H_0 zamietame práve vtedy, ak $x \in K$, t. j. v pokuse tri razy vytiahneme čiernu loptičku.

Ak naopak, z vytiahnutých bude aspoň jedna biela, tak H_0 nezamietame. Test je zostrojený a je na mieste zamyslieť sa, aké má vlastnosti. Začnime tým, že naše rozhodovanie vždy(!) sprevádzajú dve chyby. Tie chyby sú rôznej povahy, a tak dostali rôzne pomenovania – chyba prvého a chyba druhého druhu:

Chyba prvého druhu	spočíva v tom, že zamietame H_0, a pritom H_0 je platná, H_0 je objektívne správna.
Chyba druhého druhu	spočíva v tom, že nezamietneme H_0, a pritom H_0 neplatí, objektívne správnou je H_1.

Naša situácia s loptičkami naozaj dobre ilustruje, že týmto chybám sa nevyhneme! Je skutočne dôležité uvedomiť si, že takéto chyby budú sprevádzať *každý* štatistický test. Stav vecí zobrazme tabuľkou:

	Objektívny stav:	
Naše rozhodnutie:	H_0 je platná	H_0 je neplatná
zamietame H_0	chyba 1. druhu	správne rozhodnutie
nezamietame H_0	správne rozhodnutie	chyba 2. druhu

Uvažujúc o kvalite testu (založeného na oblasti K), nás samozrejme zaujíma, s akou pravdepodobnosťou sa dopúšťame možných chýb. Uvedomujeme si, že aj keď je v škatuli len jedna čierna, aj tak je možné, že H_0 zamietneme, lebo náhoda sa s nami zahrá tak, že tri razy vytiahneme tú jedinú čiernu! Aká je pravdepodobnosť takejto možnosti? Ide o pravdepodobnosť chyby prvého druhu (symbol || nahrádza slová „pritom"):

$$\begin{aligned} &P(\text{zamietame } H_0, \text{ pritom } H_0 \text{ je platná}) = P(\text{zamietame } H_0 \,||\, p=0.2) = \\ &= P((X_1,X_2,X_3)=(1,1,1) \,||\, p=0.2) = P(X_1=1, X_2=1, X_3=1 \,||\, p=0.2) = \\ &= P(X_1=1)P(X_2=1)P(X_3=1) = 0.2^3 = 0.008. \end{aligned}$$

Pravdepodobnosť chyby prvého druhu je iba 0.008, to je celkom prijateľné! Teraz počítajme pravdepodobnosť chyby druhého druhu:

$$\begin{aligned} &P(\text{nezamietame } H_0, \text{ pritom } H_0 \text{ neplatí}) = P(\text{nezamietame } H_0 \,||\, p=0.4) = \\ &= P((X_1,X_2,X_3) \neq (1,1,1) \,||\, p=0.4) = 1 - P((X_1,X_2,X_3)=(1,1,1) \,||\, p=0.4) = \\ &= 1 - P(X_1=1)P(X_2=1)P(X_3=1) = 1 - 0.4^3 = 1 - 0.064 = 0.936. \end{aligned}$$

Toto je iste šokujúce zistenie, náš test má veľkú slabinu, pretože s pravdepodobnosťou až 0.936 nezamietame neplatnú hypotézu! K tejto situácii sa vrátime v poznámke 8.1.4. Teraz sformulujeme použité pojmy v definícii 8.1.3.

8.1.3 Definícia. *Kritickou oblasťou* K rozumieme tú podmnožinu priestoru R^n, pre ktorú nastatie náhodnej udalosti $\{(X_1, X_2, \ldots, X_n) \in K\}$ má za následok zamietnutie hypotézy H_0. Hovoríme, že test hypotézy H_0 proti alternatíve H_1 je založený na kritickej oblasti K.

Chyba 1. druhu spočíva v rozhodnutí zamietnuť H_0, a pritom je H_0 objektívne platná. Pravdepodobnosť chyby 1. druhu sa nazýva *hladina* testu a označuje sa α. Teda

$$\alpha = P((X_1, X_2, \ldots, X_n) \in K \parallel H_0 \text{ je platná}).$$

Takýto stav popisujeme aj slovami – kritická oblasť K má *veľkosť* α.

Chyba 2. druhu spočíva v rozhodnutí nezamietnuť H_0, aj keď je v skutočnosti neplatná, lebo objektívne platnou je H_1. Pravdepodobnosť chyby 2. druhu sa označuje β. Teda

$$\beta = P((X_1, X_2, \ldots, X_n) \notin K \parallel H_1 \text{ je platná}).$$

Číslo $1 - \beta$ sa nazýva sila testu.

8.1.4 Poznámka. Vráťme sa k príkladu 8.1.2. Zistili sme, že test založený na K má príliš malú silu (voľne povedané, veľmi často prijímame neplatnú H_0). Preto zmeňme kritickú oblasť a odteraz zamietajme H_0, ak spomedzi troch vytiahnutých sú aspoň dve čierne. To znamená, že definujeme novú kritickú oblasť. Odteraz je ňou polpriestor L

$$L = \{(x_1, x_2, x_3) \in R^3 : x_1 + x_2 + x_3 \geq 2\}.$$

Pre test založený na kritickej oblasti L platí:

$$\alpha = P(\text{zamietame } H_0, \text{ pritom } H_0 \text{ je platná}) = P((X_1, X_2, X_3) \in L \parallel p = 0.2) =$$
$$= P(X_1 + X_2 + X_3 \geq 2 \parallel p = 0.2) = 3 \cdot 0.2^2 \cdot 0.8 + 0.2^3 = 0.096 + 0.008 = 0.104.$$

$$\beta = P(\text{nezamietame } H_0, \text{ pritom } H_0 \text{ nie je platná}) = P((X_1, X_2, X_3) \notin L \parallel p = 0.4) =$$
$$= P(X_1 + X_2 + X_3 \leq 1 \parallel p = 0.4) = 0.6^3 + 3 \cdot 0.4 \cdot 0.6^2 = 0.216 + 0.432 = 0.648.$$

Ako vidíme, znížili sme síce pravdepodobnosť chyby 2. druhu (aj keď nie dostatočne), ale dôsledkom je zvýšenie pravdepodobnosti chyby 1. druhu. Je zrejmé, že zlepšenie testu dosiahneme len tak, keď zvýšime rozsah náhodného výberu (úloha 8.5.1).

Ak rozsah náhodného výberu je daný, pevný, tak pravdepodobnosti chýb 1. druhu a 2. druhu sa chovajú ako „spojené nádoby". Zmenšenie jednej z nich má za následok zväčšenie druhej (v rámci príkladu 8.1.2, resp. tejto poznámky, zmenšenie β z 0.936 na 0.648, malo za následok nárast α z 0.008 na 0.104).

Treba sa rozhodnúť, čo zvolíme za kritérium optimality testu. V našej situácii je celkom logické, snažiť sa minimalizovať súčet pravdepodobností oboch chýb, t. j. nájsť kritickú oblasť tak, aby súčet $\alpha + \beta$ bol čo najmenší. V takom prípade test založený na kritickej oblasti L je lepší, ako test založený na oblasti K (pretože $0.104 + 0.648 < 0.008 + 0.936$).

Vezmime na vedomie, že v reálnom testovaní ide o testy, v ktorých aspoň jedna hypotéza je zloženou a ako ukážeme v ďalšom, postavenie hypotéz nie je také symetrické ako v našej (elementárnej) situácii. Preto sa prijala iná koncepcia optimality rozhodovania. Štandardný test prebieha tak, že stanovíme hranicu α pre pravdepodobnosť chyby 1. druhu a spomedzi všetkých oblastí veľkosti nanajvýš α vyberieme tú, pri ktorej má test najmenšiu pravdepodobnosť chyby 2. druhu (má teda najväčšiu silu). Takéto uvažovanie vedie k pojmu *najsilnejšej* kritickej oblasti veľkosti α, ktorý je však už mimo rámca tohoto textu.

Zmysel nasledujúceho príkladu je ukázať, ako nájdeme kritickú oblasť testu v prípade, keď hladina testu je zadaná.

8.1.5 Príklad. Nech $(X_1, X_2, \ldots, X_9)$ je náhodný výber z rozdelenia $N(\mu, 1)$. Nájdime kritickú oblasť K veľkosti $\alpha = 0.05$, pre test o strednej hodnote

$$H_0: \mu = 5 \quad \text{proti} \quad H_1: \mu = 6.$$

Inými slovami – máme určiť kritickú oblasť K tak, aby pre hladinu testu platilo $\alpha = 0.05$. Ďalej určíme silu testu, ktorý je založený na kritickej oblasti K.

Riešenie. V príklade 8.1.2 bol rozsah výberu $n = 3$, kritická oblasť K bola preto množina v R^3. Intuitívne sme pracovali s testovacou štatistikou $X_1 + X_2 + X_3$, pretože jej hodnoty predstavovali počet čiernych loptičiek medzi tromi vybratými. Teraz $n = 9$, a tak K bude podmnožinou v R^9. Akou štatistikou si pomôžeme pri hľadaní kritickej oblasti pre H_0?

Oprieme sa o výberový priemer $\bar{X}$. Vieme, že je najlepším nevychýleným odhadom strednej hodnoty, teda parametra μ. Ak je H_0 platná, hodnoty $\bar{X}$ registrujeme v okolí čísla 5. Tie hodnoty $\bar{X}$, ktoré ležia príliš napravo od čísla 5, sú zrejme kritickými pre H_0. Preto pre tvar K máme

$$K = \{(x_1, x_2, \ldots, x_9) \in R^9 : \tfrac{1}{9} \textstyle\sum_{i=1}^{9} x_i \geq k\}.$$

Treba stanoviť k tak, aby veľkosť K sa rovnala zadanej hodnote 0.05, t. j. aby hladina testu $\alpha = 0.05$. Preto má platiť

$$P((X_1, X_2, \ldots, X_9) \in K \parallel H_0 \text{ je platná}) = 0.05.$$

To znamená, že hľadáme k, aby

$$P(\tfrac{1}{9} \textstyle\sum_{i=1}^{9} X_i \geq k \parallel H_0 \text{ je platná}) = 0.05, \quad \text{teda } P(\bar{X} \geq k \parallel \mu = 5) = 0.05.$$

Ak H_0 je platná, $(X_1, X_2, \ldots, X_9)$ je náhodný výber z $N(5, 1)$, a preto $\bar{X} \sim N(5, \frac{1}{9})$. Uvedená podmienka pre k bude platiť, keď

$$P(\bar{X} < k \parallel \mu = 5) = 0.95.$$

Normovaním $\bar{X}$ dostávame

$$P\left(\frac{\bar{X} - 5}{\sqrt{\frac{1}{9}}} < \frac{k - 5}{\sqrt{\frac{1}{9}}}\right) = 0.95,$$

odkiaľ je zrejmé, že číslo $\frac{k-5}{\sqrt{\frac{1}{9}}} = 3(k - 5)$ musí byť 0.95-kvantil rozdelenia $N(0, 1)$. Z tabuliek kvantilov (str. 167), získame 0.95-kvantil, je ním číslo 1.645. Zo vzťahu

$$3(k - 5) = 1.645, \text{ dostávame } k = 5 + \frac{1.645}{3} = 5 + 0.5483 \approx 5.55.$$

Kritická oblasť K veľkosti 0.05 je stanovená:

$$K = \{(x_1, x_2, \ldots, x_9) \in R^9 : \tfrac{1}{9} \textstyle\sum_{i=1}^{9} x_i \geq 5.55\}.$$

Hodnote 5.55 hovoríme *kritická hodnota*, pretože ak sa testovacia štatistika $\bar{X}$ realizuje hodnotou väčšou, resp. rovnajúcou sa 5.55, tak H_0 zamietame. Ďalej nás zaujíma sila testu. Najprv vyčíslime pravdepodobnosť β chyby 2. druhu. Pripomeňme, že

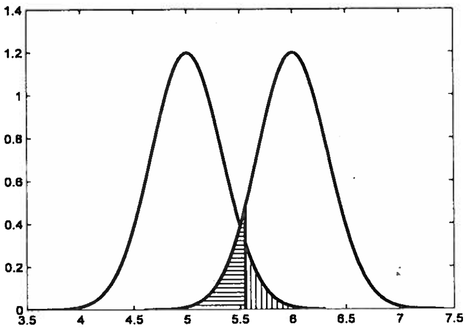
$$\beta = P((X_1, X_2, \ldots, X_9) \notin K \parallel H_1 \text{ je platná}).$$

Ak H_1 je platnou, tak $\mu = 6$ a pre výberový priemer máme $\bar{X} \sim N(6, \frac{1}{9})$. Preto platí

$$\beta = P((X_1, X_2, \dots, X_9) \notin K \parallel \mu = 6) = P(\bar{X} < 5.55 \parallel \mu = 6) = P\left(\frac{\bar{X} - 6}{\sqrt{\frac{1}{9}}} < \frac{5.55 - 6}{\sqrt{\frac{1}{9}}} \right) = F_N(-1.35) =$$

$$= 0.08851 \approx 0.09.$$

Sila testu založeného na oblasti K sa rovná $1 - 0.09 = 0.91$. Pravdepodobnosti α, β oboch chýb sú znázornené na obr. 8-1.



Obr. 8-1. Kritická hodnota $k = 5.55$. Zvislo šrafovaná plocha má obsah $\alpha = 0.05$. Vodorovne šrafovaná plocha má obsah $\beta = 0.09$ (a predstavuje pravdepodobnosť chyby 2. druhu)

8.1.6 Príklad. Na výrobe určitého výrobku sa podieľajú dve linky. Prvá z nich, označme ju A, vyrába s 5-percentnou nepodarkovosťou, kým druhá linka B má nepodarkovosť až 10-percentnú. Predstavme si, že máme pred sebou kontajner s 1000 výrobkami a vieme o ňom iba to, že buď obsahuje výrobky z linky A, alebo naopak, jeho obsah pochádza z linky B. Predpokladajme, že sme realizovali náhodný výber rozsahu $n = 100$ a na základe zistenia počtu nepodarkov medzi vybratými máme rozhodnúť, z ktorej linky tých 1000 výrobkov pochádza. Formulujme úlohu ako štatistický test na hladine $\alpha = 0.1$ a nájdime jeho silu.

Riešenie. Náhodný výber modelujeme 100-ticou nezávislých veličín X_i, $X_i \sim A(p)$, kde X_i sú indikátory nepodarku (príklad 6.2.5). Ak výrobky boli vyrobené linkou A, tak $p = 0.05$, ak výrobky pochádzajú z linky B, tak $p = 0.10$. Ide o test o parametri p alternatívneho rozdelenia:

$$H_0: p = 0.05 \quad \text{proti} \quad H_1: p = 0.10$$

Testovacou štatistikou je veličina $S = X_1 + X_2 + \dots + X_{100}$, ktorá predstavuje počet nepodarkov medzi 100 vybratými. Veličinou S budeme identifikovať kritickú oblasť K. Zrejme kritickými pre H_0 sú príliš veľké hodnoty S, a preto K má tvar

$$K = \{ (x_1, x_2, \dots, x_{100}) \in R^{100} : x_1 + x_2 + \dots + x_{100} \geq k \}.$$

V ňom treba určiť kritickú hodnotu k tak, aby veľkosť oblasti K sa rovnala 0.1. Veličina S je diskrétna, pričom $S \sim Bi(100, p)$. Z článku 5.1 vieme, že $Bi(n, p)$ môžeme aproximovať rozdelením $N(np, np(1-p))$, s ktorým sa bude pracovať pohodlnejšie. To znamená, že

ak platí H_0, tak na S nazeráme ako na veličinu s rozdelením $N(5, 4.75)$,
ak platí H_1, tak na S nazeráme ako na veličinu s rozdelením $N(10, 9)$.

Test má mať hladinu 0.1, a tak musí platiť

$$P((X_1, X_2, \dots, X_{100}) \in K \parallel H_0 \text{ je platná}) = 0.1,$$

teda

$$P(S \geq k \parallel S \sim N(5, 4.75)) = 0.1,$$

čo znamená, že má platiť

$$P(S < k \parallel S \sim N(5, 4.75)) = 0.9.$$

Normovanie S dáva

$$P\left(\frac{S-5}{\sqrt{4.75}} < \frac{k-5}{\sqrt{4.75}} \right) = 0.9,$$

t. j. $\frac{k-5}{\sqrt{4.75}}$ sa musí rovnať 0.9-kvantilu $N(0, 1)$, teda $\frac{k-5}{\sqrt{4.75}} = 1.282$. Pre k máme

$$k = 5 + 1.282\sqrt{4.75} = 7.7941 \approx 7.79.$$

Pretože S má len celočíselné hodnoty, hľadanú oblasť K definujeme vzťahom

$$K = \{ (x_1, x_2, \dots, x_{100}) \in R^{100} : x_1 + x_2 + \dots + x_{100} \geq 8 \}.$$

Ak počet nepodarkov medzi 100 vybratými sa rovná aspoň 8, H_0 zamietame. Ak počet nepodarkov je menší ako 8, H_0 nezamietame. Uvedomme si, že hladinu testu, ktorá sa mala rovnať 0.1, sme trochu porušili (binomické rozdelenie sme aproximovali normálnym rozdelením). Teraz určíme silu nášho testu. Najprv počítajme pravdepodobnosť chyby 2. druhu.

$$\beta = P(S < 8 \parallel S \sim N(10, 9)) = P\left(\frac{S - 10}{\sqrt{9}} < \frac{8 - 10}{\sqrt{9}} \right) = F_N(-0.67) = 1 - 0.74857 \approx 0.25.$$

To znamená, že sila testu založeného na kritickej oblasti K sa rovná 0.75.

8.2 Test jednoduchej hypotézy proti zloženej alternatíve

V praxi je častejšia situácia, keď aspoň jedna z hypotéz je zloženou, t. j. určuje viacprvkovú množinu hodnôt parametra, najčastejšie však nekonečnú množinu. V tomto článku sa venujeme prípadu, keď H_0 je jednoduchá a H_1 je zložená. Situácia umožňuje hovoriť o novom, dôležitom pojme – o silofunkcii testu. Začneme jednoduchým príkladom.

8.2.1 Príklad. O minci máme rozhodnúť, či je, alebo nie je falošná. Buď na nej padá znak s pravdepodobnosťou 0.5 – vtedy jej hovorme normálna, v opačnom prípade ju nazývajme falošnou. Pochopiteľne, rozhodnúť máme na základe pokusu, ktorý pozostáva napríklad zo 16 nezávislých hodov mincou. Formulujme úlohu ako štatistický test a zamyslime sa nad jeho vlastnosťami.

Riešenie. Ak veličina X_i je indikátor padnutia znaku v i-tom hode, tak za model zloženého pokusu považujeme 16-ticu $X_1, X_2, \dots, X_{16}$ nezávislých náhodných veličín, z ktorých každá má rozdelenie $A(p)$. Veličiny $X_1, X_2, \dots, X_{16}$ predstavujú náhodný výber z rozdelenia $A(p)$. Asi každý súhlasí s postojom, že mincu považujeme za normálnu, pokiaľ nezískame dostatočné argumenty v prospech tvrdenia, že je falošná. Preto test formulujeme v tvare

$$H_0: p = \tfrac{1}{2} \quad \text{proti} \quad H_1: p \neq \tfrac{1}{2}$$

Takto zvolenej alternatíve hovoríme *obojstranná*. V našej situácii je to prirodzená voľba.

Zatiaľ nepredpisujme hladinu testu, len intuitívne zvoľme kritickú oblasť K a uvažujme o vlastnostiach testu založeného na K.

Tvar kritickej oblasti je jasný. Ak je minca normálna, počet núl a počet jednotiek v realizácii $x = (x_1, x_2, \dots, x_{16})$ náhodného výberu je približne rovnaký (to znamená, že súčet $x_1 + x_2 + \dots + x_{16}$ kolíše okolo čísla 8). Kritickými pre H_0 sú tie body x, v ktorých sa počet jednotiek výrazne líši od počtu núl, t. j. súčet $x_1 + x_2 + \dots + x_{16}$ je alebo príliš malý, alebo príliš veľký. Testovacou štatistikou bude teda opäť $S = X_1 + X_2 + \dots + X_{16}$. Pre prehľadnosť uveďme hodnoty veličiny S, ktoré považujeme za málo pravdepodobné v prípade, že H_0 je platná (môžeme im hovoriť, že sú „kritické" pre H_0).

$H(S)$:	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
kritické:	0	1	2	3	4								12	13	14	15	16

Avšak slovom *kritické hodnoty štatistiky S* sa pri testovaní označujú hranice: v našom prípade by nimi boli hodnoty 4 a 12. Hodnoty štatistiky S, pri ktorých nemáme dôvod zamietať H_0 sú: 5, 6, 7, 8, 9, 10 a 11. Kritickú oblasť K v R^{16} sme intuitívne zvolili takto:

$$K = \{(x_1, x_2, \dots, x_{16}) \in R^{16} : x_1 + x_2 + \dots + x_{16} \le 4\} \cup \{(x_1, x_2, \dots, x_{16}) \in R^{16} : x_1 + x_2 + \dots + x_{16} \ge 12\}.$$

Určme hladinu testu založeného na K. Ak H_0 je platná, štatistika S má rozdelenie $\mathrm{Bi}(16, \frac{1}{2})$.

$$\alpha = P(X_1 + X_2 + \dots + X_{16} \le 4 \,||\, p = \tfrac{1}{2}) + P(X_1 + X_2 + \dots + X_{16} \ge 12 \,||\, p = \tfrac{1}{2}) =$$

$$= \textstyle\sum_{k=0}^{4} \binom{16}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{16-k} + \sum_{k=12}^{16} \binom{16}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{16-k} =$$

$$= \left(\tfrac{1}{2}\right)^{16} \left(\tbinom{16}{0} + \tbinom{16}{1} + \tbinom{16}{2} + \tbinom{16}{3} + \tbinom{16}{4} + \tbinom{16}{16} + \tbinom{16}{15} + \tbinom{16}{14} + \tbinom{16}{13} + \tbinom{16}{12} \right) =$$

$$= \left(\tfrac{1}{2}\right)^{16} 2\left(\tbinom{16}{0} + \tbinom{16}{1} + \tbinom{16}{2} + \tbinom{16}{3} + \tbinom{16}{4} \right) = \left(\tfrac{1}{2}\right)^{15} (1 + 16 + 120 + 560 + 1820) \approx 0.077.$$

To je vcelku prijateľná hodnota. S pravdepodobnosťou 0.077 zamietame H_0, aj keď je platná, t. j. s pravdepodobnosťou 0.077 tvrdíme, že minca je falošná a nemáme pravdu.

Nové v tejto situácii je to, že pravdepodobnosť chyby 2. druhu závisí od hodnoty p, ktorú deklaruje H_1. Lenže H_1 deklaruje nekonečne veľa hodnôt! To znamená, že

$$\beta = P((X_1, X_2, \dots, X_{16}) \notin K \,||\, H_1 \text{ je platná}) = P(5 \le X_1 + X_2 + \dots + X_{16} \le 11 \,||\, p \ne \tfrac{1}{2}) = \beta(p)$$

je funkcia parametra p. Hodnoty funkcie $\beta(.)$, napr. pre $p = 0.6, 0.7, 0.8, 0.9$, dostaneme buď použitím nejakého výpočtového prostredia, alebo jednoducho z tabuliek (str.168):

$$\beta(0.6) = P(5 \le X_1 + X_2 + \dots + X_{16} \le 11 \,||\, p = 0.6) = \textstyle\sum_{k=5}^{11} \binom{16}{k} 0.6^k 0.4^{16-k} = 0.8285,$$

$$\beta(0.7) = P(5 \le X_1 + X_2 + \dots + X_{16} \le 11 \,||\, p = 0.7) = \textstyle\sum_{k=5}^{11} \binom{16}{k} 0.7^k 0.3^{16-k} = 0.5498,$$

$$\beta(0.8) = P(5 \le X_1 + X_2 + \dots + X_{16} \le 11 \,||\, p = 0.8) = \textstyle\sum_{k=5}^{11} \binom{16}{k} 0.8^k 0.2^{16-k} = 0.2018,$$

$$\beta(0.9) = P(5 \le X_1 + X_2 + \dots + X_{16} \le 11 \,||\, p = 0.9) = \textstyle\sum_{k=5}^{11} \binom{16}{k} 0.9^k 0.1^{16-k} = 0.0170.$$

Ľahko sa presvedčíme, že rovnaké hodnoty dostaneme pre $p = 0.4, 0.3, 0.2, 0.1$. Je zrejmé, že test má malú silu, veď ak $p = 0.6$, tak s pravdepodobnosťou až 0.8285 nezamietame H_0, hoci nie je platná (mincu akceptujeme ako normálnu a ona je falošná!).

Ak chceme pracovať s lepším testom, je nutné zväčšiť rozsah výberu. Urobíme to v nasledujúcom príklade.

8.2.2 Príklad. Pokračujme v príklade 8.2.1. Formulácia testu ostáva, ale teraz nech $n = 50$ a využime CLV na aproximáciu rozdelenia $\mathrm{Bi}(50, p)$ (vďaka tomu už nebudeme závislí od výpočtového prostredia). Kým v 8.2.1 sme kritickú oblasť zvolili len intuitívne, teraz si kladieme za úlohu stanoviť K tak, aby mala veľkosť 0.05 – inými slovami – aby test bol testom na hladine 0.05. Nakoniec nájdime priebeh silofunkcie testu.

Riešenie. Pretože $X_i \sim A(p)$, tak $E(X_i) = p$, $var(X_i) = p(1-p)$. Odvolávajúc sa na CLV, odteraz považujeme $S = X_1 + X_2 + \dots + X_{50}$ za veličinu s normálnym rozdelením, presnejšie,

ak platí H_0, tak na S nazeráme ako na veličinu s rozdelením $N(25, 12.5)$,

ak platí H_1, tak na S nazeráme ako na veličinu s rozdelením $N(50p, 50p(1-p))$, $p \ne \frac{1}{2}$.

Pre K má platiť

$$P((X_1, X_2, \dots, X_{50}) \in K \,||\, H_0 \text{ je platná}) = 0.05,$$

pričom tvar K bude

$$K = \{(x_1, x_2, \dots, x_{50}) \in R^{50} : x_1 + x_2 + \dots + x_{50} \le k_1\} \cup \{(x_1, x_2, \dots, x_{50}) \in R^{50} : x_1 + x_2 + \dots + x_{50} \ge k_2\}.$$

Treba určiť kritické hodnoty k_1, k_2. Určíme ich z podmienok

$$P(X_1 + X_2 + \dots + X_{50} < k_1 \,||\, X_1 + X_2 + \dots + X_{50} \sim N(25, 12.5)) = 0.025,$$

$$P(X_1 + X_2 + \dots + X_{50} < k_2 \,||\, X_1 + X_2 + \dots + X_{50} \sim N(25, 12.5)) = 0.975.$$

Normovanie $S = X_1 + X_2 + \dots + X_{50}$ vedie k podmienke $P(\frac{S-25}{\sqrt{12.5}} < \frac{k_1 - 25}{\sqrt{12.5}}) = 0.025$, čo znamená, že $\frac{k_1 - 25}{\sqrt{12.5}}$ je 0.025-kvantil $N(0,1)$, t. j. číslo -1.96. Preto $k_1 = 25 - 1.96\sqrt{12.5} = 18.07$.

Z druhej podmienky máme, že $\frac{k_2 - 25}{\sqrt{12.5}}$ sa rovná 1.96, teda $k_2 = 25 + 1.96\sqrt{12.5} = 31.93$.

Vzhľadom na to, že S má v skutočnosti iba celočíselné hodnoty (veď je súčtom núl a jednotiek), stanovme K takto:

$$K = \{(x_1, x_2, \dots, x_{50}) \in R^{50} : x_1 + x_2 + \dots + x_{50} \le 18\} \cup \{(x_1, x_2, \dots, x_{50}) \in R^{50} : x_1 + x_2 + \dots + x_{50} \ge 32\}.$$

Veľkosť K sa približne rovná 0.05 (približne preto, lebo binomické rozdelenie sme aproximovali rozdelením normálnym a pre jednoduchosť – urobili sme to bez korekcie).

Nakoniec máme určiť priebeh silofunkcie testu založeného na oblasti K. Najprv určíme niekoľko hodnôt funkcie $\beta(p)$, ktorá predstavuje hodnoty pravdepodobnosti chyby 2. druhu. Začnime napr. hodnotou $\beta(p)$ pre $p = 0.6$. Parametre normálneho rozdelenia, ktorým aproximujeme binomické rozdelenie $\mathrm{Bi}(50, 0.6)$ sú

$$\mu = np = 50 \cdot 0.6 = 30, \quad \sigma^2 = np(1-p) = 50 \cdot 0.6 \cdot 0.4 = 12.$$

Dostávame (teraz použijúc korekciu – pozri príklad 5.1.8)

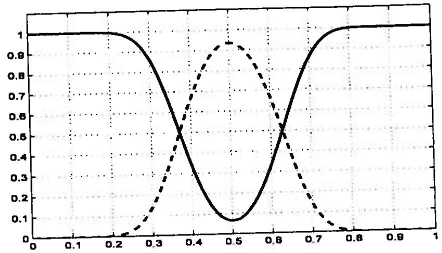
$$\beta(0.6) = P(19 \le S \le 31) = F_N\left(\frac{31.5 - 30}{\sqrt{12}}\right) - F_N\left(\frac{18.5 - 30}{\sqrt{12}}\right) = F_N(0.43) - F_N(-3.32) \approx 0.67.$$

Ak $p = 0.7$, tak parametre normálneho rozdelenia sú

$$\mu = np = 50 \cdot 0.7 = 35, \quad \sigma^2 = np(1-p) = 50 \cdot 0.7 \cdot 0.3 = 10.5,$$

a preto $\beta(0.7) = P(19 \le S \le 31) = F_N\left(\frac{31.5-35}{\sqrt{10.5}}\right) - F_N\left(\frac{18.5-35}{\sqrt{10.5}}\right) = F_N(-1.08) - F_N(-5.09) \approx 0.14.$

Analogicky určíme ostatné hodnoty $\beta(p)$. Graf funkcie $\beta(\cdot)$ je na obrázku 8-2 znázornený prerušovanou čiarou, graf silofunkcie $1-\beta(\cdot)$ plnou čiarou.



Obr. 8-2. Plnou čiarou je znázornený graf silofunkcie testu z príkladu 8.2.2

Čo vidíme z grafu $\beta(p)$? Ak minca je falošná a skutočná pravdepodobnosť padnutia znaku $p = 0.6$, tak s pravdepodobnosťou až 0.67 prijímame hypotézu, že minca je normálna. S takýmto testom sotva môžeme byť spokojní.

8.2.3 Príklad. Koľko hodov mincou je potrebných na to, aby test

$$H_0: p = \tfrac{1}{2} \quad \text{proti} \quad H_1: p \ne \tfrac{1}{2}$$

na hladine 0.05 bol taký, že ak objektívne správnou je $p = 0.6$, tak pravdepodobnosť chyby druhého druhu sa rovná (približne) 0.2?

Riešenie. Najprv nájdime kritickú oblasť K tak, aby hladina testu $\alpha = 0.05$. Pretože testovacou štatistikou je súčet $S = X_1 + X_2 + \ldots + X_n$, oblasť K veľkosti 0.05 nie je možné stanoviť bez toho, aby sme poznali n. Veď ak je H_0 platná, $S \sim \text{Bi}(n, \frac{1}{2})$ a $E(S) = \frac{n}{2}$, a tak kým nepoznáme n, nevieme ani len $E(S)$, a preto nemôžme určiť oblasť K danej veľkosti.

Úloha sa dá riešiť iteračným postupom a, samozrejme, za pomoci výpočtového prostredia. V našom prípade by sme vzali napr. $n = 100$, našli by sme K (veľkosti 0.05) a určili by sme $\beta(p; 100)$, pre $p = 0.6$. Keby bolo $\beta(0.6; 100) > 0.2$, zvýšili by sme rozsah výberu n a pre nové n opäť našli novú K (veľkosti 0.05) a určili $\beta(0.6; n)$. Po konečnom počte krokov by sme našli hľadané n. Pre zaujímavosť, hľadané $n = 192$.

8.3 Testy o strednej hodnote normálneho rozdelenia

V tomto článku pôjde o testy parametra μ v rozdelení $N(\mu, \sigma^2)$. Pritom zloženou hypotézou môže byť nielen H_1 (ako to bolo v článku 8.2), ale aj nulová hypotéza H_0. Preberieme prípad nielen obojstrannej, ale aj jednostrannej alternatívy. Pretože normálne rozdelenie má dva parametre, treba rozlišovať jednak prípad, keď druhý parameter σ poznáme a aj prípad (v praxi častejší), keď σ je neznáme. Začneme jednoduchším prípadom, keď smerodajná odchýlka σ je známa. Pretože budeme pracovať s kvantilmi normálneho rozdelenia, zaveďme označenie u_α pre α-kvantil rozdelenia $N(0,1)$. Zopakujme pojem kvantilu. Ak $X \sim N(0,1)$, $\alpha \in (0,1)$, tak α-kvantil je číslo u_α definované vzťahom:

$$P(X < u_\alpha) = \alpha, \text{ čo znamená, že pre opačnú udalosť platí: } P(X \ge u_\alpha) = 1 - \alpha.$$

8.3.1 Príklad. Nech $X \sim N(0,1)$, $\alpha \in (0,1)$. Ukážme, že pre kladné c platí

$$P(|X| \ge c) = \alpha \quad \text{práve vtedy, keď} \quad c = u_{1-\alpha/2}.$$

$$\text{To znamená, že platí: } P(|X| \ge u_{1-\alpha/2}) = \alpha.$$

Riešenie. Z aditivity máme $P(|X| \ge c) = P(X \le -c) + P(X \ge c)$. Pretože $N(0, 1)$ je symetrické rozdelenie, zrejme $P(X \le -c) = P(X \ge c)$, odkiaľ

$$P(|X| \ge c) = P(X \le -c) + P(X \ge c) = 2P(X \ge c).$$

Nakoniec, $P(|X| \ge c) = \alpha \;\Leftrightarrow\; P(X \ge c) = \alpha/2 \;\Leftrightarrow\; P(X < c) = 1 - \alpha/2 \;\Leftrightarrow\; c = u_{1-\alpha/2}$.

8.3.2 Veta. Nech $X_1, X_2, \ldots, X_n$ je náhodný výber z $N(\mu, \sigma^2)$, kde σ je známy parameter. Pre test

$$H_0: \mu = \mu_0 \quad \text{proti} \quad H_1: \mu \ne \mu_0$$

je

$$K = \{x \in R^n : |\bar{x} - \mu_0| \ge \frac{\sigma u_{1-\alpha/2}}{\sqrt{n}}\}$$

kritickou oblasťou na hladine α (teda K má veľkosť α).

Dôkaz. Tvar K je celkom prirodzený. Kritickými pre H_0 sú zrejme tie hodnoty výberového priemeru $\bar{X}$, ktoré sú významne vzdialené od μ_0. Oblasť K má veľkosť α, ak pravdepodobnosť zamietania platnej H_0 sa rovná α, teda ak $P((X_1, X_2, \ldots, X_n) \in K \,\|\, \mu = \mu_0) = \alpha$.
Z definovania K máme

$$\{(X_1, X_2, \ldots, X_n) \in K\} = \{|\bar{X} - \mu_0| \ge \frac{\sigma u_{1-\alpha/2}}{\sqrt{n}}\},$$

z čoho plynie, že treba ukázať: Ak $\mu = \mu_0$, tak $P\left(|\bar{X} - \mu_0| \ge \frac{\sigma u_{1-\alpha/2}}{\sqrt{n}}\right) = \alpha$.

Ak však $\mu = \mu_0$, tak $\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$ a normovanie $\bar{X}$ dáva: $\frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n} \sim N(0, 1)$.

Podľa príkladu 8.3.1 platí

$$P\left(\left|\frac{\bar{X} - \mu_0}{\sigma}\sqrt{n}\right| \ge u_{1-\alpha/2}\right) = \alpha, \text{ čo znamená, že } P\left(\frac{|\bar{X} - \mu_0|}{\sigma}\sqrt{n} \ge u_{1-\alpha/2}\right) = \alpha, \text{ teda}$$

$$P\left(|\bar{X} - \mu_0| \ge \frac{\sigma u_{1-\alpha/2}}{\sqrt{n}}\right) = \alpha,$$

a práve to sme mali ukázať.

8.3.3 Príklad. O smerodajnej odchýlke σ náhodných chýb meracieho prístroja vieme, že sa rovná 0.15. Systematická chyba meracieho prístroja sa eliminuje nastavením prístroja a detekuje sa meraním etalónu, pri ktorom správnou nameranou hodnotou je $\mu_0 = 10.00$. Nezávislými meraniami etalónu za rovnakých podmienok sme získali hodnoty

10.10	10.16	9.85	10.19	9.96	10.29	9.88	10.25	9.94

Je možné odchýlky od hodnoty μ_0 vysvetliť náhodnými chybami s nulovou strednou hodnotou, alebo je dôvod predpokladať, že ide o systematickú chybu, ktorú je potrebné nastavením prístroja odstrániť?
Riešenie. O náhodných chybách meracích prístrojov sa implicitne predpokladá, že majú normálne rozdelenie s nulovou strednou hodnotou. Ak stredná hodnota sa nerovná nule,

ide o systematickú chybu. Tá sa prejaví pri meraní etalónu tým, že stredná hodnota meranej veličiny sa nerovná 10. V našom prípade máme rozhodnúť, či merania etalónu dávajú dosť argumentov na zamietnutie hypotézy o tom, že stredná hodnota sa rovná 10. Použijeme test

$$H_0\colon \mu = 10 \quad \text{proti} \quad H_1\colon \mu \neq 10$$

Zvoľme hladinu 0.05 (to je najčastejšia voľba). Podľa vety 8.3.2 je kritickou oblasťou testu oblasť K

$$K = \{ x \in R^9 : |\bar{x} - \mu_0| \geq \frac{0.15\, u_{0.975}}{\sqrt{9}} \}.$$

Treba zistiť, či naša realizácia x (našich nameraných 9 hodnôt) patrí do K, alebo nepatrí do K. Výberový priemer sa realizuje hodnotou $\bar{x} = 10.0689$, to znamená, že $|\mu_0 - \bar{x}| = 0.0689$. Pretože $u_{0.975} = 1.96$, výraz na pravej strane nerovnosti sa rovná $0.15 \cdot 1.96/\sqrt{9} = 0.098$. Je zrejmé, že realizácia x náhodného výberu nie je bodom K (0.0689 nie je väčšie alebo rovné 0.098). To znamená, že na hladine 0.05 nezamietame H_0.

Na základe daných dát nezamietame hypotézu, že stredná hodnota náhodných chýb sa rovná nule. Avšak, pozor! Uvedomme si, že sme urobili rozhodnutie, lenže to, aký je skutočný stav vecí (t. j. aká je skutočná hodnota μ) – to ostáva pred nami skryté.

8.3.4 Objektívny stav prírody sa s istotou štatistickými metódami síce nedozvieme, avšak vieme, že kvalitu testu (na danej hladine) opisuje silofunkcia testu (čo je temer to isté, ako pravdepodobnosť chyby 2. druhu). Cieľom tohoto odseku je ukázať, že pre test z vety 8.3.2 vieme stanoviť silu testu. Nájdime niekoľko hodnôt funkcie $\beta(\mu)$ pre test zostrojený v 8.3.3.

Riešenie. Pravdepodobnosť chyby 2. druhu je pravdepodobnosť nezamietania H_0, keď v skutočnosti H_0 nie je správnou, v našom prípade, keď skutočné μ sa nerovná μ_0 (=10.0). Formálnejšie,

$$\beta(\mu) = P((X_1, X_2, \dots, X_9) \notin K \,\|\, \mu \neq \mu_0) = P(|\bar{X} - 10.0| < \frac{0.15\, 1.96}{\sqrt{9}} \quad \|\, \bar{X} \sim N(\mu, \frac{0.15^2}{9}))$$

Pretože $\frac{0.15\, 1.96}{3} = 0.098$ a $\sigma(\bar{X}) = \sqrt{\frac{0.15^2}{9}} = \frac{0.15}{3} = 0.05$, máme

$$\beta(10.05) = P(|\bar{X} - 10| < 0.098 \,\|\, \bar{X} \sim N(10{:}05, 0.05^2) = P(9.902 < \bar{X} < 10.098 \,\|\, \bar{X} \sim N(10.05, 0.05^2) =$$
$$= F_N\left(\frac{10.098 - 10.05}{0.05}\right) - F_N\left(\frac{9.902 - 10.05}{0.05}\right) = F_N(0.96) - F_N(-2.96) = 0.82993 \approx 0.83,$$

$$\beta(10.10) = P(|\bar{X} - 10| < 0.098 \,\|\, \bar{X} \sim N(10.10, 0.05^2) = P(9.902 < \bar{X} < 10.098 \,\|\, \bar{X} \sim N(10.10, 0.05^2) =$$
$$= F_N\left(\frac{10.098 - 10.10}{0.05}\right) - F_N\left(\frac{9.902 - 10.10}{0.05}\right) = F_N(-0.04) - F_N(-3.96) = 0.48405 \approx 0.48,$$

$$\beta(10.15) = P(|\bar{X} - 10| < 0.098 \,\|\, \bar{X} \sim N(10.15, 0.05^2) = P(9.902 < \bar{X} < 10.098 \,\|\, \bar{X} \sim N(10.15, 0.05^2) =$$
$$= F_N\left(\frac{10.098 - 10.15}{0.05}\right) - F_N\left(\frac{9.902 - 10.15}{0.05}\right) = F_N(-1.04) - F_N(-4.96) = 0.14917 \approx 0.15.$$

Analogicky dostaneme hodnoty $\beta(9.95)$, $\beta(9.90)$, $\beta(9.85)$, pričom (presvedčte sa) dostaneme identické výsledky. Získané hodnoty hovoria, že ak skutočná μ je v okolí piatich stotín deklarovanej hodnoty μ_0 (=10.0), tak pravdepodobnosť nezamietania H_0 je nepríjemne veľká (až 0.83), avšak pomerne rýchle klesá, veď pre $\mu = 10.15$ je $\beta(10.15)$ menšia ako 0.15.

8.3.5 Uvažujme o teste o strednej hodnote, v ktorom aj H_0 je zloženou, napr. o teste v tvare

$$H_0\colon \mu \leq \mu_0 \quad \text{proti} \quad H_1\colon \mu > \mu_0 .$$

Uvedomme si, že je treba nanovo definovať hladinu testu. Doteraz H_0 deklarovala jedinú hodnotu μ_0 a pravdepodobnosť chyby 1. druhu bolo číslo

$$P((X_1, X_2, \dots, X_n) \in K \,\|\, \mu = \mu_0).$$

Teraz však pravdepodobnosť chyby 1. druhu je funkcia na množine hodnôt μ, $\mu \leq \mu_0$, t. j. na intervale $(-\infty, \mu_0)$, pretože každej μ, $\mu \leq \mu_0$ môžeme priradiť $P((X_1, X_2, \dots, X_n) \in K \,\|\, \mu)$.

Hovoríme, že test

$$H_0\colon \mu \leq \mu_0 \quad \text{proti} \quad H_1\colon \mu > \mu_0$$

je testom na hladine α (inými slovami: oblasť K má veľkosť α), ak

$$\sup \{ P((X_1, X_2, \dots, X_n) \in K \,\|\, \mu) : \mu \leq \mu_0 \} = \alpha,$$

t. j. suprémum nekonečnej množiny $\{ P((X_1, X_2, \dots, X_n) \in K \,\|\, \mu) : \mu \leq \mu_0 \}$ sa rovná α.

8.3.6 **Veta.** Nech $(X_1, X_2, \dots, X_n)$ je náhodný výber z $N(\mu, \sigma^2)$, kde σ je známa hodnota. Pre test

$$H_0\colon \mu \leq \mu_0 \quad \text{proti} \quad H_1\colon \mu > \mu_0$$

je

$$K = \{ x \in R^n : \bar{x} \geq \mu_0 + \frac{\sigma\, u_{1-\alpha}}{\sqrt{n}} \}$$

kritickou oblasťou na hladine α.

Dôkaz. Tvar K je prirodzený, pretože kritickými pre H_0 sú zrejme hodnoty $\bar{X}$ ležiace príliš ďaleko napravo od μ_0. Analogicky ako v dôkaze vety 8.3.2 sa ukáže, že ak $\mu = \mu_0$, tak

$$P\left(\bar{X} - \mu_0 \geq \frac{\sigma\, u_{1-\alpha}}{\sqrt{n}}\right) = \alpha, \quad \text{teda} \quad P\left(\bar{X} \geq \mu_0 + \frac{\sigma\, u_{1-\alpha}}{\sqrt{n}}\right) = \alpha.$$

To, čo je teraz nové, je ukázať, že ak $\mu < \mu_0$, tak $P\left(\bar{X} \geq \mu_0 + \frac{\sigma\, u_{1-\alpha}}{\sqrt{n}}\right) \leq \alpha$, pričom výberový priemer $\bar{X}$ má rozdelenie $N\left(\mu, \frac{\sigma^2}{n}\right)$. To je však celkom jednoduché, stačí si načrtnúť grafy hustôt $N\left(\mu_0, \frac{\sigma^2}{n}\right)$ a $N\left(\mu, \frac{\sigma^2}{n}\right)$, kde $\mu < \mu_0$ a porovnať odpovedajúce plochy pod hustotami.

8.3.7 Samozrejme, v situáciách, keď ide o test strednej hodnoty normálneho rozdelenia, častejším je prípad, keď parameter σ nie je známy. Doteraz bolo možné pracovať s výberovým priemerom $\bar{X}$ takto:

$$\text{Ak } \mu = \mu_0, \text{ tak } \bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right), \text{ z čoho zrejme } \frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n} \sim N(0, 1).$$

Preto

$$P\left(\frac{\bar{X} - \mu_0}{\sigma}\sqrt{n} \geq u_{1-\alpha}\right) = \alpha, \text{ odkiaľ máme výsledok: } P\left(\bar{X} - \mu_0 \geq \frac{\sigma\, u_{1-\alpha}}{\sqrt{n}}\right) = \alpha.$$

To je fakticky prvá časť dôkazu 8.3.6, ktorú sme neurobili (len sa odvolali na dôkaz 8.3.2). Teraz sa venujme prípadu, keď parameter σ nie je známy.

Namiesto štatistiky $\frac{\bar{X}-\mu_0}{\sigma}\sqrt{n}$ vezmime výberovú štatistiku

$$\frac{\bar{X}-\mu_0}{S}\sqrt{n}$$

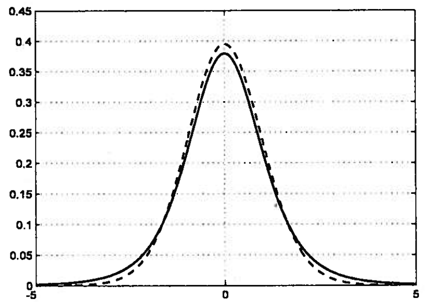
s ktorou sa dá pracovať, aj keď σ nepoznáme. Potrebujeme však zistiť, aké je jej rozdelenie.

8.3.8 Definícia. Hovoríme, že náhodná veličina T má Studentovo t-rozdelenie, $T \sim t(n)$, ak jej rozdelenie určuje hustota f daná vzťahom

$$f(x;n) = k_n\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad \text{pre všetky } x \in \mathrm{R}.$$

Parameter rozdelenia n sa nazýva *počet stupňov voľnosti*. Hodnotou parametra môže byť ľubovoľné prirodzené číslo. Skalár k_n (závisí od n) zaručuje, že integrál z hustoty (cez celú číselnú os) sa rovná jednej (čo je základná vlastnosť každej hustoty). Jeho explicitné vyjadrenie je trochu komplikované (vyžaduje znalosť gamma funkcie). Pre predstavu uveďme, že hodnoty k_n sú z intervalu (0.3, 0.4).

Z predpisu hustoty je zrejmé, že je to párna funkcia. Graf má zvonovitý tvar a nápadne pripomína graf hustoty normovaného Gaussovho rozdelenia. Dá sa ukázať, že ak $n \ge 2$, tak $E(T) = 0$ a ak $n \ge 3$, tak už existuje aj variancia rozdelenia a platí: $var(T) = n/(n-2)$.



Obr. 8-3. Graf hustôt: $t(5)$ plnou čiarou, $t(25)$ prerušovanou čiarou

To, čo budeme potrebovať, sú kvantily t-rozdelenia. Na rozdiel od kvantilov rozdelenia N(0, 1), teraz tabuľka kvantilov má dva vstupy: p a n. Preto bude označenie kvantilov komplikovanejšie. Napr. p-kvantil $t(n)$ pre $n = 10$, budeme označovať: $t_p(10)$. Tabuľky uvádzajú hodnoty kvantilov len pre najčastejšie hodnoty p, $p = 0.9$, 0.95, 0.975, 0.99, 0.995.

8.3.9 Veta. Ak $(X_1, X_2, \ldots, X_n)$ je náhodný výber z rozdelenia N(μ, σ²), tak výberová štatistika

$$\frac{\bar{X}-\mu}{S}\sqrt{n}$$

má Studentovo t-rozdelenie s $n-1$ stupňami voľnosti.

Dôkaz vety je nad naše možnosti. Veta však umožňuje zostrojiť test o strednej hodnote normálneho rozdelenia. Všimnime si, že výsledky, ktoré uvádza veta 8.3.10, sú dokonalé analógie viet 8.3.2 a 8.3.6.

8.3.10 Veta. Nech $(X_1, X_2, \ldots, X_n)$ je náhodný výber z rozdelenia N(μ, σ²), kde σ² je neznámy parameter.

a) Pre test $\quad H_0: \mu = \mu_0 \quad$ proti $\quad H_1: \mu \ne \mu_0$

je

$$K = \{x \in R^n : |\bar{x} - \mu_0| \ge \frac{s\,t_{1-\alpha/2}(n-1)}{\sqrt{n}}\}$$

kritická oblasť veľkosti α. To znamená, že test založený na K je testom na hladine α.

b) Pre test $\quad H_0: \mu \le \mu_0 \quad$ proti $\quad H_1: \mu > \mu_0$

je

$$K = \{x \in R^n : \bar{x} \ge \mu_0 + \frac{s\,t_{1-\alpha}(n-1)}{\sqrt{n}}\}$$

kritická oblasť veľkosti α. Teda test založený na oblasti K je testom na hladine α.

c) Pre test $\quad H_0: \mu \ge \mu_0 \quad$ proti $\quad H_1: \mu < \mu_0$

je

$$K = \{x \in R^n : \bar{x} \le \mu_0 - \frac{s\,t_{1-\alpha}(n-1)}{\sqrt{n}}\}$$

kritická oblasť veľkosti α. Test založený na oblasti K je testom na hladine α.

Dôkaz vety sa urobí presne tak, ako sme urobili dôkaz viet 8.3.2 a 8.3.6. Zásadný rozdiel je v tom, že kým v prípade, keď poznáme σ, sme dokázali určiť aj pravdepodobnosť chyby 2. druhu, teraz musíme priznať, že to nedokážeme (taká úloha presahuje rámec tohoto textu).

To, že nevieme stanoviť silu testu, teda že máme pod kontrolou iba pravdepodobnosť chyby 1. druhu a nemáme predstavu o veľkosti pravdepodobnosti chyby 2. druhu, má zásadný význam v reálnej situácii pri formulovaní testu. Ilustrujme to príkladom.

8.3.11 Príklad. Medzi základné parametre automobilového motora patrí spotreba a množstvo uvoľnených emisií. Uvažujme o spotrebe. Predstavme si, že chceme deklarovať, že spotreba konkrétneho motora nepresiahne hodnotu $\mu_0 = 5$ [litrov/100 km]. Za rovnakých podmienok urobíme napr. 16 pokusov tak, aby sme mohli predpokladať nezávislosť jednotlivých opakovaní pokusu. Tu sú dáta:

5.3 4.9 4.5 5.3 4.7 4.3 4.7 5.1 4.4 5.0 3.9 4.5 4.6 4.5 5.5 4.6

- Prvá formulácia testu

$$H_0: \mu \le 5.0 \quad \text{proti} \quad H_1: \mu > 5.0$$

na hladine 0.05 je kritickou oblasťou

$$K = \{x \in R^n : \bar{x} \ge \mu_0 + \frac{s\,t_{1-\alpha}(n-1)}{\sqrt{n}}\} = \{x \in R^{16} : \bar{x} \ge 5 + \frac{s\,t_{0.95}(15)}{\sqrt{16}}\}$$

Z realizácie výberu máme: $\bar{x} = 4.7375$, $s = 0.4209$, $t_{0.95}(15) = 1.7531$. Pretože 4.7375 nie je väčšie ako $5 + \frac{s\,t_{0.95}(15)}{\sqrt{16}} = 5 + \frac{0.4209\ 1.7531}{4} = 5 + 0.1845 = 5.1845$, H_0 nezamietame.

Môžeme byť spokojní, lebo obrazne povedané, naše dáta H_0 nezamietli. Avšak naozaj môžeme byť spokojní? Veď nevieme, s akou pravdepodobnosťou sme chybili!

Keď však (vo všeobecnosti) v nejakom teste (ako takom) H_0 *zamietame*, tak vieme s akou pravdepodobnosťou robíme chybu – tá pravdepodobnosť sa predsa rovná hladine testu. *Preto skutočnosť, ktorú chceme testovaním potvrdiť, dajme do formulácie alternatívy* a dúfajme, že dáta nám umožnia H_0 zamietnuť. V našom príklade:

- Nová formulácia testu

$$H_0: \mu \geq 5.0 \quad \text{proti} \quad H_1: \mu < 5.0$$

Kritickou oblasťou veľkosti 0.05 je $K = \{x \in R^{16} : \bar{x} \leq 5 - \frac{s\, t_{0.95}(15)}{\sqrt{16}}\}$. Zisťujeme, že naša realizácia patrí do K, pretože $\bar{x} = 4.7375$ je menej ako $5 - 0.1845 = 4.8155$. H_0 zamietame. Prijímame teda hypotézu, že spotreba je menšia ako 5 [litrov/100 km]. Pritom vieme, že pravdepodobnosť toho, že robíme chybné rozhodnutie je iba 0.05.

Zopakujme, čo sme už dva razy povedali: Majme vždy na zreteli, že stav prírody, t. j. objektívne platnú skutočnosť, sa prostredníctvom štatistického testu nikdy nedozvieme.

8.4 Testy o stredných hodnotách dvoch normálne rozdelených veličín

V praktických situáciách často chceme porovnať účinky dvoch postupov, resp. efektívnosť dvoch metód, či porovnať rýchlosť dvoch programov, alebo spotrebu dvoch motorov a pod. Porovnanie chceme urobiť na základe dát, ktoré predstavujú realizácie náhodných výberov z jedného a druhého základného súboru. Predpokladajme, že rozdelenia oboch základných súborov sú normálne a ide nám o porovnanie ich stredných hodnôt.

8.4.1 Nech $(X_1, X_2, \dots, X_m)$ je náhodný výber z rozdelenia $N(\mu_1, \sigma^2)$ a $(Y_1, Y_2, \dots, Y_n)$ je náhodný výber z rozdelenia $N(\mu_2, \sigma^2)$, pričom veličiny $X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_n$ sú nezávislé. Potom sú nezávislé aj veličiny $\bar{X}$ a $\bar{Y}$ a ak položíme $Z = \bar{X} - \bar{Y}$, tak platí

$$E(Z) = \mu_1 - \mu_2, \quad var(Z) = \frac{\sigma^2}{m} + \frac{\sigma^2}{n} = \sigma^2\left(\frac{m+n}{mn}\right).$$

Normovanie Z dáva W

$$W = \frac{Z - E(Z)}{\sqrt{var(Z)}} = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{m+n}{mn}}} = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma}\sqrt{\frac{mn}{m+n}}, \quad W \sim N(0, 1).$$

Ak parameter σ poznáme, W je výberová štatistika (jej hodnoty vieme vyčísliť z výberov) a môže hrať úlohu testovacej štatistiky pre testy

$$H_0: \mu_1 = \mu_2 \quad \text{proti} \quad H_1: \mu_1 \neq \mu_2, \quad \text{resp.} \quad H_0: \mu_1 \leq \mu_2 \quad \text{proti} \quad H_1: \mu_1 > \mu_2.$$

Kritické oblasti týchto testov určíme podobne ako v 8.3.2, resp. 8.3.6 (úloha 8.5.11). Nasledujúci odsek je venovaný prípadu – v praxi častejšiemu – keď parameter σ nepoznáme.

8.4.2 Nech platí to, čo uvádza odsek 8.4.1, teraz však predpokladajme, že σ nepoznáme. V takom prípade veličina W nie je výberová štatistika (lebo ju nevieme vyčísliť z výberu). Analogicky ako v 8.3.7 je však prirodzené, neznámy parameter σ odhadnúť. Pretože oba výbery, t. j. $(X_1, X_2, \dots, X_m)$, aj $(Y_1, Y_2, \dots, Y_n)$ sú výbermi z rozdelení s neznámou, ale *rovnakou* varianciou, použijeme všetkých týchto $m+n$ veličín na odhad σ^2.

Označme

$$S_X^2 = \frac{1}{m-1}\sum_{i=1}^m (X_i - \bar{X})^2, \quad S_Y^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2.$$

Zrejme $\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{i=1}^n (Y_i - \bar{Y})^2 = (m-1)S_X^2 + (n-1)S_Y^2$ a ak tento súčet $m+n$ kvadrátov predelíme číslom $(m+n-2)$, tak štatistika

$$S_p^2 = \frac{1}{m+n-2}\left((m-1)S_X^2 + (n-1)S_Y^2\right)$$

je nevychýleným odhadom σ^2, t.j. $E(S_p^2) = \sigma^2$ (čo vôbec nie je problém ukázať, pozri úlohu 8.5.12). Všimnime si, že veta 8.4.3 je obdoba vety 8.3.9.

8.4.3 Veta. Nech $(X_1, X_2, \dots, X_m)$ je náhodný výber z rozdelenia $N(\mu_1, \sigma^2)$ a $(Y_1, Y_2, \dots, Y_n)$ je náhodný výber z rozdelenia $N(\mu_2, \sigma^2)$, a pritom veličiny $X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_n$ sú nezávislé. Nech platia označenia z 8.4.2, pričom

$$S_p = \sqrt{\frac{1}{m+n-2}\left((m-1)S_X^2 + (n-1)S_Y^2\right)}\,.$$

Potom náhodná veličina

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p}\sqrt{\frac{mn}{m+n}}$$

má Studentovo t-rozdelenie s $m+n-2$ stupňami voľnosti, teda $T \sim t(m+n-2)$.

Aj keď dôkaz vety je nad naše sily, všimnime si analógiu so štatistikou W z 8.4.1. Formálny rozdiel medzi W a T je len v tom, že namiesto parametra σ (smerodajnej odchýlke oboch rozdelení) v štatistike T vystupuje jej odhad S_p. Význam vety je zrejmý. V prípade, že variancia základných súborov (ktorých stredné hodnoty chceme porovnať) je rovnaká, štatistika T môže byť použitá na konštrukciu kritickej oblasti pre test o stredných hodnotách. Pritom postup odvodenia je rovnaký ako v článku 8.3.

8.4.4 Veta. Nech platia predpoklady a označenia z odseku 8.4.2 a vety 8.4.3.

a) Pre test $\quad H_0: \mu_1 = \mu_2 \quad \text{proti} \quad H_1: \mu_1 \neq \mu_2$
je

$$K = \{(x, y) \in R^{m+n} : |\bar{x} - \bar{y}| \geq \frac{s_p\, t_{1-\alpha/2}(m+n-2)}{\sqrt{\frac{mn}{m+n}}}\}$$

kritickou oblasťou veľkosti α.

b) Pre test $\quad H_0: \mu_1 \leq \mu_2 \quad \text{proti} \quad H_1: \mu_1 > \mu_2$
je

$$K = \{(x, y) \in R^{m+n} : \bar{x} - \bar{y} \geq \frac{s_p\, t_{1-\alpha}(m+n-2)}{\sqrt{\frac{mn}{m+n}}}\}$$

kritickou oblasťou veľkosti α.

8.4.5 Príklad. Cieľom experimentu je potvrdiť, že nový typ nárazníka (typ B) je vhodnejší, ako pôvodný (typ A). Účinnosť nárazníka sa meria cenou, ktorú je treba zaplatiť na opravu auta, ktoré v rýchlosti 10 km/hod. čelne narazí do prekážky. Dvadsať nových áut (konkrétnej značky a toho istého modelu) sme náhodne rozdelili na dve skupiny. Autám prvej skupiny boli namotnované nárazníky typu A, autám druhej skupiny nárazníky typu B. Dáta $x_1, x_2, \dots, x_{10}$ predstavujú ceny opráv áut po náraze v prvej skupine a $y_1, y_2, \dots, y_{10}$ sú ceny opráv áut v druhej skupine. Z dát sme dostali realizácie výberových štatistík

$$\bar{x} = 198, \; s_x = 18.25, \; \text{resp.} \; \bar{y} = 176, \; s_y = 15.88.$$

Na hladine 0.05 realizujme test o stredných hodnotách vo verzii (b) z vety 8.4.4

$$\mathrm{H}_0\colon \mu_1 \le \mu_2 \quad \text{proti} \quad \mathrm{H}_1\colon \mu_1 > \mu_2$$

Treba rozhodnúť, či 20-tica $(x, y) = (x_1, x_2, \dots, x_{10}, y_1, y_2, \dots, y_{10})$ patrí, alebo nepatrí do kritickej oblasti K definovanej vo vete 8.4.4. Pre realizáciu štatistiky S_p máme

$$s_p = \sqrt{\frac{1}{18}\,(9s_x^2 + 9s_y^2)} = \sqrt{\frac{1}{18}(2997.5625 + 2269.5696)} = \sqrt{292.61845} = 17.1061,$$

nakoniec

$$\frac{s_p\, t_{1-\alpha}(m+n-2)}{\sqrt{\frac{mn}{m+n}}} = \frac{17.1061\; 1.7341}{\sqrt{\frac{100}{20}}} = \frac{29.6637}{2.2361} = 13.27$$

Keďže $\bar{x} - \bar{y} = 198 - 176 = 22$, tak je zrejmé, že (x, y) patrí do K, a preto na hladine 0.05 zamietame H_0. Prijímame hypotézu, ktorá tvrdí, že nový typ, typ B, je účinnejší ako typ A.

8.4.6 Poznámka. O t-teste (test z viet 8.3.10, resp. 8.4.4) je známe, že je *robustný* v tom zmysle, že predpoklad normality môže byť do istej miery porušený a napriek tomu test má vlastnosti aké má, keď predpoklady sú bezo zvyšku splnené. V teste z vety 8.4.4 okrem normality základných súborov vystupuje ešte predpoklad o tom, že oba výbery pochádzajú zo súborov, ktoré majú *rovnakú varianciu*. Toto je chúlostivý bod tohoto testu, pretože nemáme prostriedky na to, aby sme overili jeho splnenie. V mnohých knižkách sa odporúča najprv previesť test o tom, že máme do činenia s normálne rozdelenými súbormi s *rovnakou varianciou* (Fisherovým–Snedecorovým F-testom) a keď hypotézu o rovnosti variancií nezamietneme, až potom pokračovať v testovaní stredných hodnôt podľa 8.4.4.

Dobre, predstavme si, že F-test urobíme a v ňom nezamietneme hypotézu o rovnakých varianciách. Vari sme tým dokázali, že súbory majú rovnakú varianciu? Samozrejme, že nie. Testom sa nikdy nič nedokazuje – testom sa len zamieta, resp. nezamieta nulová hypotéza (na danej hladine). Keď naše rozhodovanie je založené na nezamietaní hypotéz v dvoch po sebe idúcich testoch, prestáva byť jasné, aké sú pravdepodobnosti možných chýb. Táto problematika však svojou zložitosťou prekračuje rámec tohoto textu. Kompetentné informácie je možné nájsť v [1], článok 8.1.

Štatistická komunita už medzičasom našla spôsoby, ako testovať stredné hodnoty *bez predpokladu rovnosti variancií*. Sformulujeme postup testovania podľa Satterthwaite 1946. Za testovaciu štatistiku vezmeme štatistiku T z vety 8.4.4, ale počet stupňov voľnosti df (degrees of freedom) nebude určený jednoduchým súčtom $m + n - 2$, ale dostaneme ho zaokrúhlením (na celé číslo) výsledku, ktorý vznikne vyčíslením výrazu

$$\frac{[(S_X^2/m) + (S_Y^2/n)]^2}{\frac{(S_X^2/m)^2}{m-1} + \frac{(S_Y^2/n)^2}{n-1}}$$

Takáto modifikácia testu z vety 8.4.4 bude testom, ktorého hladina sa približne rovná α.

8.4.7 Príklad. Učiteľ vytvoril informatickú pomôcku (IP) a pripravil experiment, aby na reálnych výsledkoch overil jej vklad do vyučovacieho procesu. Skupinu všetkých študentov rozdelil na dve časti A, B tak, aby z hľadiska všeobecnej disponovanosti študentov vznikli dve rovnocenné skupiny. Kým skupina A prešla štandardným vyučovacím procesom bez použitia IP, skupina B využívala IP. Na skúške, v ktorej mali študenti preukázať úroveň vedomostí z danej oblasti, boli dosiahnuté takéto výsledky:

$$\begin{array}{llll} \text{Skupina A:} & m = 14, & \bar{x} = 79.5, & s_x = 6.36, \\ \text{Skupina B:} & n = 15, & \bar{y} = 85.9, & s_y = 8.22. \end{array}$$

Vzhľadom na okolnosti, formulujme test takto:

$$\mathrm{H}_0\colon \mu_1 \ge \mu_2 \quad \text{proti} \quad \mathrm{H}_1\colon \mu_1 < \mu_2$$

(očakávajúc, že test zamietne H_0). Test je modifikáciou 8.4.4 (b), a preto pre kritickú oblasť testu máme

$$\mathrm{K} = \Big\{ (x, y) \in R^{m+n}\colon \bar{y} - \bar{x} \ge \frac{s_p\, t_{1-\alpha}(df)}{\sqrt{\frac{mn}{m+n}}} \Big\} = \Big\{ (x, y) \in R^{29}\colon \bar{y} - \bar{x} \ge \frac{s_p\, t_{0.95}(df)}{\sqrt{\frac{210}{14+15}}} \Big\}$$

kde df dostaneme tak, ako je uvedené v 8.4.6.

$$\frac{[(S_X^2/m) + (S_Y^2/n)]^2}{\frac{(S_X^2/m)^2}{m-1} + \frac{(S_Y^2/n)^2}{n-1}} = \frac{\left[\frac{6.36^2}{14} + \frac{8.22^2}{15}\right]^2}{\frac{(40.45/14)^2}{13} + \frac{(67.57/15)^2}{14}} = \frac{(2.889 + 4.505)^2}{0.642 + 1.449} = \frac{54.66}{2.091} = 26.14, \; \text{t. j.} \; df = 26.$$

$$s_p = \sqrt{\frac{1}{m+n-2}\left((m-1)s_X^2 + (n-1)s_Y^2\right)} = \sqrt{\frac{1}{27}(13 \cdot 40.45 + 14 \cdot 67.57)} = \sqrt{54.5122} = 7.38,$$

nakoniec

$$\bar{y} - \bar{x} = 85.9 - 79.5 = 6.4, \quad \text{kým} \quad \frac{s_p\, t_{1-\alpha}(df)}{\sqrt{\frac{mn}{m+n}}} = \frac{7.38\; 1.7056}{\sqrt{7.2414}} = \frac{12.587}{2.691} = 4.68.$$

To znamená, že realizácia (x, y) patrí do kritickej oblasti K, a preto H_0 zamietame. Prijímame hytotézu H_1, ktorá deklaruje, že IP zvýšila účinnosť vyučovacieho procesu.

8.4.8 V tomto článku sme zatiaľ hovorili o *dvojvýberových* testoch, pretože východiskom našich úvah boli dva výbery: $(X_1, X_2, \dots, X_m)$ z rozdelenia $\mathrm{N}(\mu_1, \sigma^2)$ a $(Y_1, Y_2, \dots, Y_n)$ z rozdelenia $\mathrm{N}(\mu_2, \sigma^2)$, pritom veličiny $X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_n$ boli $(m+n)$-ticou nezávislých veličín. V nasledujúcom odseku hovoríme o situácii, keď pôjde o test stredných hodnôt veličín X a Y, ktoré môžu byť závislé. Test sa bude opierať o jeden náhodný výber, ale bude to výber z *dvojrozmerného* normálneho rozdelenia.

8.4.9 Párový t-test. Uvažujme o testoch stredných hodnôt μ_1, μ_2 dvoch náhodných veličín X, Y, ktoré sa vzťahujú na ten istý prvok (ten istý objekt) základného súboru:

$$\mathrm{H}_0\colon \mu_1 = \mu_2 \quad \text{proti} \quad \mathrm{H}_1\colon \mu_1 \ne \mu_2 \qquad \text{resp.} \quad \mathrm{H}_0\colon \mu_1 \le \mu_2 \quad \text{proti} \quad \mathrm{H}_1\colon \mu_1 > \mu_2$$

(v takej situácii sú často X a Y závislé práve preto, že sa týkajú toho istého objektu). Pre konkrétnu predstavu, vezmime situáciu, keď veličina X predstavuje hodnotu cholesterolu muža pred diétou a Y hodnotu cholesterolu *toho istého* muža po diéte, napr. po 3 mesiace trvajúcej riadenej úprave jedálnička. Do experimentu je zapojených 20 mužov (nefajčiarov) vo veku medzi 40 a 50 rokov, ktorých BMI sa nachádza v intervale $\langle 25, 28 \rangle$ a majú približne rovnakú pohybovú aktivitu.

Pretože tých 20 mužov bolo vybratých náhodne, máme k dispozícii náhodný výber rozsahu 20 z dvojrozmerného normálneho rozdelenia: (X_1, Y_1), (X_2, Y_2), ... , (X_{20}, Y_{20}). Testom chceme preukázať významný vplyv diéty na hladinu cholesterolu. Pretože naše úvahy sa opierajú o jeden náhodný výber, ide o jednovýberový test. Hovorí sa mu párový, pretože dáta tvoria páry – dvojice (x_i, y_i) – v uvedenej situácii x_i je úroveň cholesterolu i-tého muža pred diétou, y_i je úroveň cholesterolu toho istého muža po diéte.

Majme teda náhodný výber (X_1, Y_1), (X_2, Y_2), ... , (X_n, Y_n) z dvojrozmerného normálneho rozdelenia. Nech $\mu_1 = \mathrm{E}(X)$, $\mu_2 = \mathrm{E}(Y)$. Čo vieme o n-tici Z_i veličín $Z_i = X_i - Y_i$, $i = 1, 2, ..., n$?

1. Veličiny Z_i sú n-ticou nezávislých veličín. To je dôsledok toho, že n-tica vektorov (X_1, Y_1), (X_2, Y_2), ... , (X_n, Y_n) je náhodným výberom z dvojrozmerného normálneho rozdelenia.
2. Z_i majú normálne rozdelenie, pričom platí:
$\mathrm{E}(Z_i) = \mathrm{E}(X_i - Y_i) = \mu_1 - \mu_2$, $var(Z_i) = var(X_i - Y_i) = var(X_i) + var(Y_i) - 2cov(X_i, Y_i) = \sigma^2$.

Uvedomme si, že neznáme hodnoty $var(X_i)$, $var(Y_i)$, $cov(X_i, Y_i)$ nezávisia na i, a preto neznáma $var(Z_i)$ je rovnaká pre všetky $i = 1, 2, ..., n$. Označili sme ju σ^2.

Veličiny $Z_1, Z_2, ..., Z_n$ predstavujú náhodný výber z rozdelenia $\mathrm{N}(\mu, \sigma^2)$, kde $\mu = \mu_1 - \mu_2$ (samozrejme, μ, μ_1, μ_2 sú neznáme) a môžeme sformulovať testy o strednej hodnote, napr.

$$\mathrm{H_0}\colon \mu_1 - \mu_2 = 0 \quad \text{proti} \quad \mathrm{H_1}\colon \mu_1 - \mu_2 \neq 0, \quad \text{to znamená:} \quad \mathrm{H_0}\colon \mu_1 = \mu_2 \quad \text{proti} \quad \mathrm{H_1}\colon \mu_1 \neq \mu_2,$$

alebo

$$\mathrm{H_0}\colon \mu_1 - \mu_2 = d \quad \text{proti} \quad \mathrm{H_1}\colon \mu_1 - \mu_2 \neq d, \quad \text{resp.} \quad \mathrm{H_0}\colon \mu_1 - \mu_2 \leq d \quad \text{proti} \quad \mathrm{H_1}\colon \mu_1 - \mu_2 > d,$$

kde d je nejaká deklarovaná hodnota (pozri príklad 8.4.10). Kritickú oblasť týchto testov konštruujeme úplne rovnako ako vo vete 8.3.10, pretože vlastne ide o jednovýberový t-test o strednej hodnote μ v rozdelení $\mathrm{N}(\mu, \sigma^2)$, keď σ nepoznáme (neznáma $\mu = \mu_1 - \mu_2$ a úlohu μ_0 má deklarovaná hodnota d). Taký test ilustruje nasledujúci príklad.

8.4.10 Príklad. Okolnosti experimentu, v ktorom išlo o preukázanie významnosti úpravy jedálnička na zníženie hladiny cholesterolu sú opísané v predchádzajúcom odseku. Teraz realizujme test na základe dát z experimentu. Pripomíname, že hodnota x_i je úroveň cholesterolu i-tého muža pred diétou a y_i je úroveň cholesterolu toho istého muža po diéte. Naše dáta – to je 20 párov, dvojíc (x_i, y_i):

(6.6, 6.3), (6.2, 5.8), (6.4, 5.8), (6.3, 5.7), ..., (6.4, 5.5), (6.2, 5.6). Prehľadnejšie snáď takto:

(x_i) :
6.6, 6.2, 6.4, 6.3, 6.5, 6.3, 6.5, 6.5, 6.7, 6.4, 6.0, 6.2, 6.7, 6.2, 6.6, 6.4, 6.7, 6.0, 6.4, 6.2

(y_i) :
6.3, 5.8, 5.8, 5.7, 5.4, 5.9, 6.0, 5.6, 5.8, 5.3, 5.3, 5.7, 6.0, 6.1, 5.5, 5.8, 5.7, 5.0, 5.5, 5.6

Na hladine 0.05 testujme

$$\mathrm{H_0}\colon \mu_1 - \mu_2 \leq 0.5 \quad \text{proti} \quad \mathrm{H_1}\colon \mu_1 - \mu_2 > 0.5$$

Keby dáta umožnili zamietnuť $\mathrm{H_0}$, prijali by sme $\mathrm{H_1}$, ktorá deklaruje: $\mu_2 < \mu_1 - 0.5$, t. j. že diéta znížila hladinu cholesterolu o viac ako 0.5 [mmol/l]. Podľa odseku 8.4.9 za realizáciu náhodného výberu z rozdelenia $\mathrm{N}(\mu, \sigma^2)$, $\mu = \mu_1 - \mu_2$, považujeme rozdiely $z_i = x_i - y_i$, (z_i):
0.3, 0.4, 0.6, 0.6, 1.1, 0.4, 0.5, 0.9, 0.9, 1.1, 0.7, 0.5, 0.7, 0.1, 1.1, 0.6, 1.0, 1.0, 0.9, 0.6

Z tejto realizácie náhodného výberu pre odhady strednej hodnoty a smerodajnej odchýlky veličiny Z máme: $\bar{z} = 0.70$, $s_z = 0.29$.

Podľa vety 8.3.10 (b), v ktorej $\mu = \mu_1 - \mu_2$, $\mu_0 = 0.5$, na hladine 0.05 je kritickou oblasťou

$$\mathrm{K} = \{ z \in R^n : \bar{z} \geq \mu_0 + \frac{s_z\, t_{1-\alpha}(n-1)}{\sqrt{n}} \} = \{ z \in R^{20} : \bar{z} \geq 0.5 + \frac{0.29\; 1.7291}{\sqrt{20}} \} = \{ z \in R^{20} : \bar{z} \geq 0.6121 \}.$$

Vidíme, že realizácia (z_i) je bodom K, a teda na hladine 0.05 zamietame $\mathrm{H_0}$. Prijímame hypotézu $\mathrm{H_1}$: $\mu_1 - \mu_2 > 0.5$, to znamená hypotézu, ktorá deklaruje, že hladina cholesterolu po diéte klesla o viac ako 0.5 [mmol/l].

8.5 Úlohy

8.5.1 V poznámke 8.1.4 sme objasnili, že test z príkladu 8.1.2 nemá dobré vlastnosti, lebo rozsah výberu je veľmi malý. Ako sme videli, ani zmenou kritickej oblasti sme nič vyhovujúce nedosiahli. Uvažujme opäť test z 8.1.2, ale teraz nech rozsah výberu $n = 20$. Vezmime za kritickú oblasť množinu $\mathrm{K} = \{x \in R^{20} : x_1 + x_2 + \ldots + x_{20} \geq 7\}$.

a) Použijúc tabuľky binomického rozdelenia (pozri tabuľky, str.170), určte jednak hladinu testu založeného na kritickej oblasti K a jednak silu takého testu.

b) Aproximujte testovaciu štatistiku $X_1 + X_2 + \ldots + X_{20}$ normálne rozdelenou veličinou Z podľa CLV a overte, že touto technikou dostanete temer rovnaké výsledky ako v časti (a), hoci podmienka $np(1-p) > 9$ splnená nie je.

8.5.2 V škatuli je 10 loptičiek, a pritom buď 4 sú biele a 6 je čiernych, alebo je to práve naopak. Predpokladajme, že môžeme vykonať náhodný výber s vrátením rozsahu 5 a rozhodneme sa, že akonáhle aspoň tri z vytiahnutých budú biele, zamietame hypotézu, že sú tam 4 biele, a teda prijímame hypotézu, že počet bielych sa rovná 6. Sformulujte úlohu ako štatistický test, určte jeho hladinu a jeho silu.

8.5.3 Nech $X_1, X_2, \ldots, X_8$ je náhodný výber z Poissonovho rozdelenia. Uvažujme o teste

$$\mathrm{H_0}\colon \lambda = 0.5 \quad \text{proti} \quad \mathrm{H_1}\colon \lambda = 1.$$

Vezmime za kritickú oblasť množinu $\mathrm{K} = \{x \in R^8 : x_1 + x_2 + \ldots + x_8 \geq 8\}$. Nájdite hladinu a silu testu založeného na kritickej oblasti K (pozri tabuľky, str. 171).

8.5.4 Nech $X_1, X_2, \ldots, X_{25}$ je náhodný výber z rozdelenia $\mathrm{N}(\mu, 16)$. Uvažujme o teste

$$\mathrm{H_0}\colon \mu = 10 \quad \text{proti} \quad \mathrm{H_1}\colon \mu = 12.$$

Určte kritickú oblasť K tak, aby hladina testu bola rovná 0.05. Nájdite silu testu založeného na oblasti K. Preveďte test, keď výberový priemer sa realizoval hodnotou 11.23.

8.5.5 Uvažujme o príklade 8.2.1, v ktorom sme o minci rozhodovali, či je normálna, alebo falošná na základe 16-tich hodov. Test bol sformulovaný takto:

$$\mathrm{H_0}\colon p = \tfrac{1}{2} \quad \text{proti} \quad \mathrm{H_1}\colon p \neq \tfrac{1}{2}$$

(kde p je pravdepodobnosť padnutia znaku). Teraz nech $n = 20$ a hypotézu $\mathrm{H_0}$ zamietajme, keď rozdiel počtu padnutých znakov a padnutých čísel je väčší alebo rovný 6. Sformulujte kritickú oblasť pomocou štatistiky $X_1 + X_2 + \ldots + X_{20}$, ktorá odpovedá uvedenému postupu a určte hladinu testu. Nájdite pravdepodobnosť chyby 2. druhu v prípade, keď skutočnou hodnotou parametra p je hodnota 0.6 (použite tabuľky rozdelenia Bi(20, p), str. 170).

8.5.6 Na základe 60-tich hodov kockou máme rozhodnúť o tom, či číslo 6 padá na kocke s pravdepodobnosťou 1/6. Nech X_i je indikátor čísla 6 v i-tom hode. Formulujte úlohu ako štatistický test s obojstrannou alternatívou a za kritickú oblasť vezmime také body R^{60}, v ktorých súčet $x_1 + x_2 + \dots + x_{60}$ je alebo menší ako 6, alebo väčší ako 14. Určte hladinu takéhoto testu využívajúc CLV.

8.5.7 Nech $X_1, X_2, \dots, X_{25}$ je náhodný výber z rozdelenia N(μ, 16). Uvažujme o teste

$$\mathrm{H_0}\colon \mu = 10 \quad \text{proti} \quad \mathrm{H_1}\colon \mu \neq 10.$$

Určte kritickú oblasť K veľkosti 0.05 (teda hladina testu sa má rovnať 0.05) a nájdite silu testu založeného na K, keď skutočná hodnota $\mu = 12$.

8.5.8 Nech $X_1, X_2, \dots, X_{25}$ je náhodný výber z rozdelenia N(μ, 16). Uvažujme o teste

$$\mathrm{H_0}\colon \mu \leq 10 \quad \text{proti} \quad \mathrm{H_1}\colon \mu > 10.$$

Určte kritickú oblasť K veľkosti 0.05 (t. j. hladina testu sa má rovnať 0.05) a nájdite silu testu založeného na K, keď skutočná hodnota $\mu = 12$. Porovnajte nájdenú hodnotu sily tohoto testu so silou v teste z úlohy 8.5.7.

8.5.9 Automat zabezpečuje váženie a balenie 250 gramových balíčkov mletej kávy. Na overenie správnosti jeho nastavenia sme náhodne vybrali 10 balení a presným vážením sme dostali hmotnosti (v gramoch)

248.2 253.4 255.3 253.6 252.7 253.3 254.1 246.2 249.4 249.2

Na hladine 0.05 preveďte test hypotézy o tom, že funkcia váženia pracuje bezchybne.

8.5.10 Výrobca deklaruje, že stredná hodnota bezporuchovej práce výrobkov istého typu je aspoň 1500 hodín. Náhodne sme vybrali 12 výrobkov tohoto typu a odmerali časy ich bezporuchovej práce:

1372 1521 1415 1613 1436 1542 1518 1426 1326 1495 1419 1549

Predpokladajme, že veličinu X – dĺžka bezporuchovej práce výrobku, môžeme považovať za veličinu s normálnym rozdelením.

Na posúdenie vierohodnosti výrobcu môžeme sformulovať testy o strednej hodnote μ buď v tvare

(1) $\mathrm{H_0}\colon \mu \geq 1500 \quad \text{proti} \quad \mathrm{H_1}\colon \mu < 1500,$

resp. v podobe

(2) $\mathrm{H_0}\colon \mu \leq 1500 \quad \text{proti} \quad \mathrm{H_1}\colon \mu > 1500.$

Preveďte oba testy na hladine $\alpha = 0.05$ a vyjadrite sa k výsledkom (interpretujte výsledky).

8.5.11 Nech $(X_1, X_2, \dots, X_m)$ je náhodný výber z rozdelenia N(μ_1, σ^2) a $(Y_1, Y_2, \dots, Y_n)$ je náhodný výber z rozdelenia N(μ_2, σ^2), pričom veličiny $X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_n$ sú nezávislé. Predpokladajme, že parameter σ^2 je známy a parametre μ_1, μ_2 sú neznáme.

(a) Pre test

$$\mathrm{H_0}\colon \mu_1 = \mu_2 \quad \text{proti} \quad \mathrm{H_1}\colon \mu_1 \neq \mu_2$$

nájdite kritickú oblasť K veľkosti α (t.j. určte K tak, aby test založený na K, bol testom na hladine α).

(b) Pre test

$$\mathrm{H_0}\colon \mu_1 \leq \mu_2 \quad \text{proti} \quad \mathrm{H_1}\colon \mu_1 > \mu_2$$

nájdite kritickú oblasť veľkosti α.

8.5.12 Nech platia predpoklady a označenia z odseku 8.4.2. Ukážte, že štatistika

$$S_p^2 = \frac{1}{m+n-2}\left((m-1)S_X^2 + (n-1)S_Y^2\right)$$

je nevychýlený odhad σ^2.

8.5.13 V poznámke 8.4.6 bolo objasnené, prečo v teste o stredných hodnotách dvoch normálnych základných súborov dávame prednosť Satterthwaitovej modifikácii dvojvýberového t-testu. Nasledujúci príklad ilustruje, že keď variancie normálnych súborov sú rovnaké, počet stupňov voľnosti df počítaný podľa výrazu z 8.4.6 sa (približne) rovná $m + n - 2$. Dáta (x_i), (y_i) sú umelo generované z normálnych súborov N(6, 0.25), resp. z N(5, 0.25).

(x_i) 5.13 5.66 6.07 6.19 6.85 5.79 6.37 6.47 6.56 5.41 6.13 6.14 5.40 5.56

(y_i) 5.10 5.88 5.19 5.79 4.93 5.54 4.46 5.24 5.74 5.33 6.24

Preveďte test podľa vety 8.4.4 vo verzii (b) na hladine 0.05. Porovnajte df počítané podľa Satterthwaitovho vzťahu s hodnotou $m + n - 2 = 14 + 11 - 2 = 23$. Samozrejme, očakávame zamietnutie hypotézy o rovnosti stredných hodnôt.

8.5.14 Cieľom experimentu bolo potvrdiť, že zmenou teploty o 40 °C sa pevnosť vlákna v ťahu významne zmení (zníži). Pri teplote 10 °C sa vykonalo 14 meraní $x_1, x_2, \dots, x_{14}$ a pri teplote 50 °C opäť 14 meraní $y_1, y_2, \dots, y_{14}$ [cN/tex] (nezávislých od tých, pri teplote 10 °C)

(x_i) 59.9 59.7 59.2 60.5 61.5 59.5 59.2 59.8 60.7 59.1 58.8 59.9 59.9 60.5

(y_i) 57.8 59.8 57.9 57.5 57.3 56.4 58.6 56.9 58.0 57.2 58.6 58.7 57.8 55.5

Sformulujte test o strednej hodnote tak, aby sme podporili hypotézu, že pevnosť vlákna zvýšením teploty o 40 °C poklesne.

8.5.15 Efektívnosť dvoch kompilátorov bola testovaná na sade 10 programov a cieľom experimentu bolo podporiť tvrdenie o tom, že druhý (nový) kompilátor pracuje rýchlejšie ako prvý. Získané dáta predstavujú časy v sekundách potrebné na kompiláciu jednotlivých programov.

Prvý kompilátor, (x_i):	3.68	4.43	4.80	3.54	2.65	3.59	4.24	4.04	3.45	3.76
Druhý kompilátor, (y_i):	3.72	4.22	4.56	3.40	2.49	3.37	4.15	4.10	3.51	3.62

Sformulujte a vykonajte test o stredných hodnotách tak, aby sme (pokiaľ možno) získali čo najviac informácie na podporu tvrdenia o tom, že druhý kompilátor pracuje rýchlejšie.

ŠTATISTICKÉ TABUĽKY

Tabuľka hodnôt distribučnej funkcie normálneho rozdelenia N(0, 1)

$$F_N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

[x	$F_N(x)$]						
0.00	0.50000	0.40	0.65542	0.80	0.78814	1.20	0.88493
0.01	0.50399	0.41	0.65910	0.81	0.79103	1.21	0.88686
0.02	0.50798	0.42	0.66276	0.82	0.79389	1.22	0.88877
0.03	0.51197	0.43	0.66640	0.83	0.79673	1.23	0.89065
0.04	0.51595	0.44	0.67003	0.84	0.79955	1.24	0.89251
0.05	0.51994	0.45	0.67364	0.85	0.80234	1.25	0.89435
0.06	0.52392	0.46	0.67724	0.86	0.80511	1.26	0.89617
0.07	0.52790	0.47	0.68082	0.87	0.80785	1.27	0.89796
0.08	0.53188	0.48	0.68439	0.88	0.81057	1.28	0.89973
0.09	0.53586	0.49	0.68793	0.89	0.81327	1.29	0.90147
0.10	0.53983	0.50	0.69146	0.90	0.81594	1.30	0.90320
0.11	0.54380	0.51	0.69497	0.91	0.81859	1.31	0.90490
0.12	0.54776	0.52	0.69847	0.92	0.82121	1.32	0.90658
0.13	0.55172	0.53	0.70194	0.93	0.82381	1.33	0.90824
0.14	0.55567	0.54	0.70540	0.94	0.82639	1.34	0.90988
0.15	0.55962	0.55	0.70884	0.95	0.82894	1.35	0.91149
0.16	0.56356	0.56	0.71226	0.96	0.83147	1.36	0.91309
0.17	0.56749	0.57	0.71566	0.97	0.83398	1.37	0.91466
0.18	0.57142	0.58	0.71904	0.98	0.83646	1.38	0.91621
0.19	0.57535	0.59	0.72240	0.99	0.83891	1.39	0.91774
0.20	0.57926	0.60	0.72575	1.00	0.84134	1.40	0.91924
0.21	0.58317	0.61	0.72907	1.01	0.84375	1.41	0.92073
0.22	0.58706	0.62	0.73237	1.02	0.84614	1.42	0.92220
0.23	0.59095	0.63	0.73565	1.03	0.84849	1.43	0.92364
0.24	0.59483	0.64	0.73891	1.04	0.85083	1.44	0.92507
0.25	0.59871	0.65	0.74215	1.05	0.85314	1.45	0.92647
0.26	0.60257	0.66	0.74537	1.06	0.85543	1.46	0.92785
0.27	0.60642	0.67	0.74857	1.07	0.85769	1.47	0.92922
0.28	0.61026	0.68	0.75175	1.08	0.85993	1.48	0.93056
0.29	0.61409	0.69	0.75490	1.09	0.86214	1.49	0.93189
0.30	0.61791	0.70	0.75804	1.10	0.86433	1.50	0.93319
0.31	0.62172	0.71	0.76115	1.11	0.86650	1.51	0.93448
0.32	0.62552	0.72	0.76424	1.12	0.86864	1.52	0.93574
0.33	0.62930	0.73	0.76730	1.13	0.87076	1.53	0.93699
0.34	0.63307	0.74	0.77035	1.14	0.87286	1.54	0.93822
0.35	0.63683	0.75	0.77337	1.15	0.87493	1.55	0.93943
0.36	0.64058	0.76	0.77637	1.16	0.87698	1.56	0.94062
0.37	0.64431	0.77	0.77935	1.17	0.87900	1.57	0.94179
0.38	0.64803	0.78	0.78230	1.18	0.88100	1.58	0.94295
0.39	0.65173	0.79	0.78524	1.19	0.88298	1.59	0.94408

1.60	0.94520	2.00	0.97725	2.40	0.99180	2.80	0.99744
1.61	0.94630	2.01	0.97778	2.41	0.99202	2.81	0.99752
1.62	0.94738	2.02	0.97831	2.42	0.99224	2.82	0.99760
1.63	0.94845	2.03	0.97882	2.43	0.99245	2.83	0.99767
1.64	0.94950	2.04	0.97932	2.44	0.99266	2.84	0.99774
1.65	0.95053	2.05	0.97982	2.45	0.99286	2.85	0.99781
1.66	0.95154	2.06	0.98030	2.46	0.99305	2.86	0.99788
1.67	0.95254	2.07	0.98077	2.47	0.99324	2.87	0.99795
1.68	0.95352	2.08	0.98124	2.48	0.99343	2.88	0.99801
1.69	0.95449	2.09	0.98169	2.49	0.99361	2.89	0.99807
1.70	0.95543	2.10	0.98214	2.50	0.99379	2.90	0.99813
1.71	0.95637	2.11	0.98257	2.51	0.99396	2.91	0.99819
1.72	0.95728	2.12	0.98300	2.52	0.99413	2.92	0.99825
1.73	0.95818	2.13	0.98341	2.53	0.99430	2.93	0.99831
1.74	0.95907	2.14	0.98382	2.54	0.99446	2.94	0.99836
1.75	0.95994	2.15	0.98422	2.55	0.99461	2.95	0.99841
1.76	0.96080	2.16	0.98461	2.56	0.99477	2.96	0.99846
1.77	0.96164	2.17	0.98500	2.57	0.99492	2.97	0.99851
1.78	0.96246	2.18	0.98537	2.58	0.99506	2.98	0.99856
1.79	0.96327	2.19	0.98574	2.59	0.99520	2.99	0.99861
1.80	0.96407	2.20	0.98610	2.60	0.99534	3.00	0.99865
1.81	0.96485	2.21	0.98645	2.61	0.99547	3.01	0.99869
1.82	0.96562	2.22	0.98679	2.62	0.99560	3.02	0.99874
1.83	0.96638	2.23	0.98713	2.63	0.99573	3.03	0.99878
1.84	0.96712	2.24	0.98745	2.64	0.99585	3.04	0.99882
1.85	0.96784	2.25	0.98778	2.65	0.99598	3.05	0.99886
1.86	0.96856	2.26	0.98809	2.66	0.99609	3.06	0.99889
1.87	0.96926	2.27	0.98840	2.67	0.99621	3.07	0.99893
1.88	0.96995	2.28	0.98870	2.68	0.99632	3.08	0.99896
1.89	0.97062	2.29	0.98899	2.69	0.99643	3.09	0.99900
1.90	0.97128	2.30	0.98928	2.70	0.99653	3.10	0.99903
1.91	0.97193	2.31	0.98956	2.71	0.99664	3.11	0.99906
1.92	0.97257	2.32	0.98983	2.72	0.99674	3.12	0.99910
1.93	0.97320	2.33	0.99010	2.73	0.99683	3.13	0.99913
1.94	0.97381	2.34	0.99036	2.74	0.99693	3.14	0.99916
1.95	0.97441	2.35	0.99061	2.75	0.99702	3.15	0.99918
1.96	0.97500	2.36	0.99086	2.76	0.99711	3.16	0.99921
1.97	0.97558	2.37	0.99111	2.77	0.99720	3.17	0.99924
1.98	0.97615	2.38	0.99134	2.78	0.99728	3.18	0.99926
1.99	0.97670	2.39	0.99158	2.79	0.99736	3.19	0.99929

Tabuľka kvantilov rozdelenia N(0, 1)

$X \sim \mathrm{N}(0,1)$, $\mathrm{P}(X < u_\alpha) = \alpha$, t. j. $F_N(u_\alpha) = \alpha$

α	0.80	0.85	0.90	0.95	0.975	0.990	0.995	0.9975
u_α	0.842	1.036	1.282	1.645	1.960	2.326	2.576	2.807

Hodnoty distribučnej funkcie F rozdelenia Bi(16, p) je možné použiť na overenie medzivýsledkov riešeného príkladu 8.2.1. Avšak pozor, hodnoty sú generované MATLABom a v ňom je distribučná funkcia implementovaná podľa vzťahu: $F(k) = \mathrm{P}(X \le k)$.

Nech $X \sim \mathrm{Bi}(16, p)$, t. j. $n = 16$ (táto hodnota sa nemení). Tabuľka uvádza hodnoty distribučnej funkcie $F = F(k, 16, p)$ rozdelenia Bi(16, p) pre jednotlivé hodnoty $p = 0.1, 0.2, \ldots, 0.9$ a pre $k = 0, 1, 2, \ldots, 16$. Tabuľka umožňuje vyčísliť napr. $\mathrm{P}(5 \le X \le 11, p = 0.6)$ takto:

$$\mathrm{P}(5 \le X \le 11, p = 0.6) = F(11, 16, 0.6) - F(4, 16, 0.6) = 0.8334 - 0.0049 = 0.8285.$$

k	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$
0	0.1853	0.0281	0.0033	0.0003	0.0000
1	0.5147	0.1407	0.0261	0.0033	0.0003
2	0.7892	0.3518	0.0994	0.0183	0.0021
3	0.9316	0.5981	0.2459	0.0651	0.0106
4	0.9830	0.7982	0.4499	0.1666	0.0384
5	0.9967	0.9183	0.6598	0.3288	0.1051
6	0.9995	0.9733	0.8247	0.5272	0.2272
7	0.9999	0.9930	0.9256	0.7161	0.4018
8	1.0000	0.9985	0.9743	0.8577	0.5982
9	1.0000	0.9998	0.9929	0.9417	0.7728
10	1.0000	1.0000	0.9984	0.9809	0.8949
11	1.0000	1.0000	0.9997	0.9951	0.9616
12	1.0000	1.0000	1.0000	0.9991	0.9894
13	1.0000	1.0000	1.0000	0.9999	0.9979
14	1.0000	1.0000	1.0000	1.0000	0.9997
15	1.0000	1.0000	1.0000	1.0000	1.0000
16	1.0000	1.0000	1.0000	1.0000	1.0000

k	$p = 0.6$	$p = 0.7$	$p = 0.8$	$p = 0.9$
0	0.0000	0.0000	0.0000	0.0000
1	0.0000	0.0000	0.0000	0.0000
2	0.0001	0.0000	0.0000	0.0000
3	0.0009	0.0000	0.0000	0.0000
4	0.0049	0.0003	0.0000	0.0000
5	0.0191	0.0016	0.0000	0.0000
6	0.0583	0.0071	0.0002	0.0000
7	0.1423	0.0257	0.0015	0.0000
8	0.2839	0.0744	0.0070	0.0001
9	0.4728	0.1753	0.0267	0.0005
10	0.6712	0.3402	0.0817	0.0033
11	0.8334	0.5501	0.2018	0.0170
12	0.9349	0.7541	0.4019	0.0684
13	0.9817	0.9006	0.6482	0.2108
14	0.9967	0.9739	0.8593	0.4853
15	0.9997	0.9967	0.9719	0.8147
16	1.0000	1.0000	1.0000	1.0000

Tabuľka kvantilov Studentovho rozdelenia s n stupňami voľnosti

n	$p = 0.90$	$p = 0.95$	$p = 0.975$	$p = 0.99$	$p = 0.995$
2	1.8856	2.9200	4.3027	6.9646	9.9248
3	1.6377	2.3534	3.1824	4.5407	5.8409
4	1.5332	2.1318	2.7764	3.7469	4.6041
5	1.4759	2.0150	2.5706	3.3649	4.0321
6	1.4398	1.9432	2.4469	3.1427	3.7074
7	1.4149	1.8946	2.3646	2.9980	3.4995
8	1.3968	1.8595	2.3060	2.8965	3.3554
9	1.3830	1.8331	2.2622	2.8214	3.2498
10	1.3722	1.8125	2.2281	2.7638	3.1693
11	1.3634	1.7959	2.2010	2.7181	3.1058
12	1.3562	1.7823	2.1788	2.6810	3.0545
13	1.3502	1.7709	2.1604	2.6503	3.0123
14	1.3450	1.7613	2.1448	2.6245	2.9768
15	1.3406	1.7531	2.1314	2.6025	2.9467
16	1.3368	1.7459	2.1199	2.5835	2.9208
17	1.3334	1.7396	2.1098	2.5669	2.8982
18	1.3304	1.7341	2.1009	2.5524	2.8784
19	1.3277	1.7291	2.0930	2.5395	2.8609
20	1.3253	1.7247	2.0860	2.5280	2.8453
21	1.3232	1.7207	2.0796	2.5176	2.8314
22	1.3212	1.7171	2.0739	2.5083	2.8188
23	1.3195	1.7139	2.0687	2.4999	2.8073
24	1.3178	1.7109	2.0639	2.4922	2.7969
25	1.3163	1.7081	2.0595	2.4851	2.7874
26	1.3150	1.7056	2.0555	2.4786	2.7787
27	1.3137	1.7033	2.0518	2.4727	2.7707
28	1.3125	1.7011	2.0484	2.4671	2.7633
29	1.3114	1.6991	2.0452	2.4620	2.7564
30	1.3104	1.6973	2.0423	2.4573	2.7500
31	1.3095	1.6955	2.0395	2.4528	2.7440
32	1.3086	1.6939	2.0369	2.4487	2.7385
33	1.3077	1.6924	2.0345	2.4448	2.7333
34	1.3070	1.6909	2.0322	2.4411	2.7284
35	1.3062	1.6896	2.0301	2.4377	2.7238
36	1.3055	1.6883	2.0281	2.4345	2.7195
37	1.3049	1.6871	2.0262	2.4314	2.7154
38	1.3042	1.6860	2.0244	2.4286	2.7116
39	1.3036	1.6849	2.0227	2.4258	2.7079
40	1.3031	1.6839	2.0211	2.4233	2.7045

Napríklad, ak $X \sim t(10)$, tak 0.95-kvantil sa rovná číslu 1.8125, t. j. $\mathrm{P}(X < 1.8125) = 0.95$.

Hodnoty distribučnej funkcie F rozdelenia Bi(20, p) je možné použiť na riešenie úlohy 8.5.1 a úlohy 8.5.5. Opäť hodnoty sú generované MATLABom, a preto $F(k) = P(X \le k)$.

Nech $X \sim \mathrm{Bi}(20, p)$. Hodnoty distribučnej funkcie $F = F(k, 20, p)$ rozdelenia Bi(20, p) pre hodnoty $p = 0.2$, 0.4, 0.5 a 0.6, v argumentoch k, pre $k = 0, 1, 2, \dots, 20$ umožňujú vyčísliť napr. $P(X \ge 7, p = 0.2)$ takto:

$$P(X \ge 7, p = 0.2) = 1 - F(6,\ 20,\ 0.2) = 1 - 0.9133 = 0.0867 \approx 0.087$$

k	$p = 0.2$	$p = 0.4$	$p = 0.5$	$p = 0.6$
0	0.0115	0.0000	0.0000	0.0000
1	0.0692	0.0005	0.0000	0.0000
2	0.2061	0.0036	0.0002	0.0000
3	0.4114	0.0160	0.0013	0.0000
4	0.6296	0.0510	0.0059	0.0003
5	0.8042	0.1256	0.0207	0.0016
6	0.9133	0.2500	0.0577	0.0065
7	0.9679	0.4159	0.1316	0.0210
8	0.9900	0.5956	0.2517	0.0565
9	0.9974	0.7553	0.4119	0.1275
10	0.9994	0.8725	0.5881	0.2447
11	0.9999	0.9435	0.7483	0.4044
12	1.0000	0.9790	0.8684	0.5841
13	1.0000	0.9935	0.9423	0.7500
14	1.0000	0.9984	0.9793	0.8744
15	1.0000	0.9997	0.9941	0.9490
16	1.0000	1.0000	0.9987	0.9840
17	1.0000	1.0000	0.9998	0.9964
18	1.0000	1.0000	1.0000	0.9995
19	1.0000	1.0000	1.0000	1.0000
20	1.0000	1.0000	1.0000	1.0000

Úloha 8.5.1. Pri stanovení sily, potrebujeme určiť $P(X \le 6, p = 0.4)$. To je hodnota $F(6, 20, 0.4)$, t. j. hodnota 0.25 (siedma hodnota v druhom stĺpci).

Úloha 8.5.5. Pri stanovení hladiny testu ide o pravdepodobnosť

$P(X \le 7, p = 0.5) + P(X \ge 13, p = 0.5) =$
$= 1 - [\,F(12, 20, 0.5) - F(7, 20, 0.5)\,] = 1 - [\,0.8684 - 0.1316\,] = 0.2632.$

Pri stanovení $\beta(0.6)$ ide o $P(8 \le X \le 12, p = 0.6) =$
$= F(12, 20, 0.6) - F(7, 20, 0.6) = 0.5841 - 0.0210 = 0.5631.$

Toto je tabuľková podpora k úlohe 8.5.3. Nasledujúce hodnoty sú hodnotami pravdepodobnostnej funkcie Poissonovho rozdelenia, pre hodnoty parametra $\lambda = 4$, resp. $\lambda = 8$ (iba tieto hodnoty sú potrebné pre riešenie úlohy 8.5.3). Ide o test

$$H_0\colon \lambda = 0.5 \quad \text{proti} \quad H_1\colon \lambda = 1$$

na základe náhodného výberu $(X_1, X_2, \dots, X_8)$.

Testovacou štatistikou je súčet $S = X_1 + \dots + X_8$. Pretože X_i sú nezávislé, a pritom všetky majú Poissonovo rozdelenie, aj veličina S má Poissonovo rozdelenie (veta 3.3.7). Kľúčom k riešeniu úlohy je uvedomiť si, že

ak platí H_0, tak štatistika S má rozdelenie Po(4), ak platí H_1, tak S má rozdelenie Po(8).

Stĺpce tabuľky sú hodnoty funkcie $e^{-\lambda}\frac{\lambda^k}{k!}$ pre $\lambda = 4$, resp. $\lambda = 8$.

k	$\lambda = 4$	$\lambda = 8$
0	0.0183	0.0003
1	0.0733	0.0027
2	0.1465	0.0107
3	0.1954	0.0286
4	0.1954	0.0573
5	0.1563	0.0916
6	0.1042	0.1221
7	0.0595	0.1396
8	0.0298	0.1396
9	0.0132	0.1241
10	0.0053	0.0993
11	0.0019	0.0722
12	0.0006	0.0481
13	0.0002	0.0296
14	0.0001	0.0169
15	0	0.009
16	0	0.0045
17	0	0.0021
18	0	0.0009
19	0	0.0004
20	0	0.0002
21	0	0.0001
22	0	0

Úlohou je nájsť hladinu testu založeného na kritickej množine
$K = \{x \in R^8 \colon x_1 + x_2 + \dots + x_8 \ge 8\}$, teda určiť

$$P(X_1 + X_2 + \dots + X_8 \ge 8 \,\|\, X_i \sim \mathrm{Po}(\lambda), \lambda = 0.5), \text{ čo je to isté, ako } P(S \ge 8 \,\|\, S \sim \mathrm{Po}(4)).$$

Ďalej treba určiť pravdepodobnosť chyby 2. druhu, resp. silu testu, to znamená určiť

$\beta = P(X_1 + X_2 + \dots + X_8 \le 7 \,\|\, X_i \sim \mathrm{Po}(\lambda), \lambda = 1)$, čo je to isté, ako $P(S \le 7 \,\|\, S \sim \mathrm{Po}(8))$. Sila testu sa rovná $1 - \beta$.

Výsledky úloh

Kapitola 1

1.1.1 a) $A \cap B' \cap C'$ b) $A \cap B \cap C'$ c) $A \cap B \cap C$
d) $(A \cap B' \cap C') \cup (A' \cap B \cap C') \cup (A' \cap B' \cap C)$
e) $A \cup B \cup C$
f) $(A \cap B \cap C') \cup (A \cap B' \cap C) \cup (A' \cap B \cap C)$
g) $A' \cap B' \cap C'$

1.1.2 Výsledky úlohy sa dajú ľahko nájsť v zadaní úlohy 1.1.3.

1.1.3 Ukážme (pre ilustráciu) vyjadrenie napr. M_3, C_6 a Z_2.
$M_3 = \{ (3, j, k) : j, k \in \{1, 2, \dots, 6\} \}$, M_3 má 36 prvkov,
$C_6 = \{ (i, 6, k) : i, k \in \{1, 2, \dots, 6\} \}$, C_6 má 36 prvkov,
$Z_2 = \{ (i, j, 2) : i, j \in \{1, 2, \dots, 6\} \}$, Z_2 má 36 prvkov.

a) 24, $P(A) = 24/216$ b) 27, $P(B) = 27/216$ c) 6, $P(C) = 6/216$
d) 91, $P(D) = 91/216$ e) 16, $P(E) = 16/216$

1.1.4 Napr. $B = A_1 \cap A_2' \cap A_3$, $C = (A_1 \cap A_2 \cap A_3') \cup (A_1 \cap A_2' \cap A_3) \cup (A_1' \cap A_2 \cap A_3)$

1.1.5 Napr. B = { (Z, Č, Z) } je elementárna udalosť, tvorí ju jeden výsledok, P(B) = 1/8,
C = { (Z, Z, Č), (Z, Č, Z), (Č, Z, Z) }, udalosť C tvoria tri výsledky pokusu, P(C) = 3/8

1.1.6 Odpovede nájdete vo formulácii úlohy 1.1.7

1.1.7 a) C = { (0, 1, 0, 0, 1) }, teda C pozostáva len z jedného stavu, P(C) = 1/32
b) D pozostáva zo šiestich stavov, P(D) = 6/32
c) E pozostáva z dvanástich stavov, P(E) = 12/32

1.2.1	a) 1/2	b) 1/16	c) 1/4	d) 1/16	e) 6/16
1.2.2	a) 0.310563	b) 0.431337	c) 0.209840	d) 0.951740	
1.2.3	a) 0.280437	b) 0.407087	c) 0.232621	d) 0.920145	
1.2.4	a) $5.950\ 10^{-5}$	b) $3.968\ 10^{-4}$	c) $1.286\ 10^{-4}$		
1.2.5	a) 2/11	b) 4/11	c) 17/55		
1.2.6	a) 14/165	b) 20/165	c) 81/165	d) 111/165	
1.2.7	a) 7/12	b) 5/6	c) 3/4		
1.2.8	a) 14/27	b) 22/27	c) 61/81		
1.2.9	20/32				
1.2.10	a) $2.40\ 10^{-4}$	b) $3.99\ 10^{-2}$	c) $2.26\ 10^{-2}$	d) 0.3262	

1.2.11 a) $1/V(9,2) = 1/72$ b) $1/V(9,4)$ c) 0.367879

1.2.12 a) 0.00421 b) 0.1921 c) 0.0213

1.2.13 $\frac{\binom{m}{i}\binom{n-m}{k-i}}{\binom{n}{k}}$

1.2.14 a) $\frac{(n-1)^n}{n^n} = (1-\frac{1}{n})^n$ b) $\frac{\binom{n}{2}(n-1)!}{n^n}$

1.2.15 $P(A_i) = \frac{\binom{n-1}{k-1}}{\binom{n}{k}} = \frac{k}{n}$ $P(B_{ij}) = \frac{\binom{n-2}{k-2}}{\binom{n}{k}} = \frac{k(k-1)}{n(n-1)}$ $P(C_{ij}) = \frac{\binom{n-2}{k-1}}{\binom{n}{k}} = \frac{k(n-k)}{n(n-1)}$

1.3.1 P(A) = 27/216, P(B) = 25/216, (súčet 11 je pravdepodobnejší)

1.3.2 P(A) = 720/46656, P(B) = 729/46656.

1.3.3 Ω = { (Z,Z), (C,Z,Z), (C,C,Z,Z), (Z,C,Z,Z), (C,C,C,Z,Z), (C,Z,C,Z,Z), (Z,C,C,Z,Z), ... }
P(A) = 0.625, P(B) = 0.859375

1.3.4 Ω = {(6), (1,6), (2,6), .., (5,6), (1,1,6), (1,2,6), (1,3,6), .., (5,5,6), (1,1,1,6), (1,1,2,6), ...}
a) A = { $(i, j, 6)$: $i, j \in \{4, 5\}$ }, A má 4 prvky, P(A) = 4/216
b) B = { $(i, j, k, 6)$: $i, j, k \in \{2, 4\}$ }, B má 8 prvkov, P(B) = 8/1296
c) C = { (6), (1,1, 6), (1, 2, 6), .., (5,5, 6), (1,1,1,1, 6), (1,1,1,2, 6), ..., (5, 5, 5, 5, 6), ...}
P(C) = 6/11

1.5.1 a) P(A) = 1/6 P(A | D) = 1/5 b) P(B) = 1/2 P(B | D) = 6/15
c) P(B) = 1/2 P(B | C) = 1/2 d) P(C) = 1/3 P(C | D) = 1/15

1.5.2 P(A) = 27/216 P(B) = 66/216 P(A | B) = 1/11 P(B | A) = 6/27

1.5.3 a) 1/5 b) 2/75 c) 17/75 d) 41/75

1.5.4 38/75

1.5.5 a) $P(B_1 | B_2) = 3/4$ b) $P(C_1 | B_2) = 1/4$
c) $P(B_1 | C_2) = 6/11$ d) $P(C_1 | C_2) = 5/11$

1.5.6 a) P(B) = 0.4, P(C) = 0.31, P(M) = 0.29 b) 0.75
c) 0.3226 d) 0.3448

1.5.7 a) 2.7% b) 0.2963

1.5.8 a) 52/70 b) 49/52

1.5.9 $P(K_1) = 0.7938$ $P(K_2) = 0.2062$

1.5.10 a) 0.097 b) 0.510

1.5.11 a) 0.02 b) $P(R_1 | H) = 0.30$, $P(R_2 | H) = 0.25$, $P(R_3 | H) = 0.45$

1.5.12 $\frac{p\,0.99}{p\,0.99+(1-p)0.02}$ Napr. ak $p = 0.07$, tak ak test bol pozitívny, tak vírus je prítomný s pravdepodobnosťou 0.7884

1.5.13 2/3

1.5.14 a) $P(H_1 | C) = 0.4$, $P(H_2 | C) = 0.3$, $P(H_3 | C) = 0.2$, $P(H_4 | C) = 0.1$, $P(H_5 | C) = 0$
b) $P(H_1 | C_1 \cap C_2) = 0.6$, $P(H_2 | C_1 \cap C_2) = 0.3$, $P(H_3 | C_1 \cap C_2) = 0.1$, $P(H_4 | C_1 \cap C_2) = 0$

1.6.1 0.9664

1.6.2 $p_1 p_2 p_3 p_4$

1.6.3 $p_1(2-p_1)\,p_2(2-p_2)\,p_3(2-p_3)\,p_4(2-p_4)$

1.6.4 a) $1-(1-p_1)(1-p_2)(1-p_3)(1-p_4)$ b) $n = 5$

1.6.5 0.8407

1.6.6 a) 0.036 b) 0.1880 c) 0.452 d) 0.1915 e) 0.5044

1.6.7 a) 0.00173 b) 0.9983 c) 0.9603

1.6.8 a) $(p\,0.001 + (1-p)\,0.855).100$ b) $(p\,0.009 + (1-p)\,0.045).100$
c) $(p\,0.099 + (1-p)\,0.095).100$ d) $(p\,0.891 + (1-p)\,0.005).100$

1.6.9 2,3% (presne 2,272%)

1.7.1 a) P(A) = 0.5, P(B) = 1/3, P(C) = 0.5, $P(A \cap B) = 1/6$, $P(A \cap B \cap C) = 1/12$
b) A, B, C sú po dvoch nezávislé c) áno, sú ako trojica nezávislé

1.7.2 a) P(A) = 1/4, P(B) = 7/12, P(C) = 7/12, $P(A \cap B) = 7/48$, $P(A \cap B \cap C) = 49/576$
b) A, B, C sú po dvoch nezávislé c) áno, sú ako trojica nezávislé

1.7.3 a) 0.0746 b) 0.1665 c) 0.1596 d) 0.00076 e) 0.26604

1.7.4 a) 0.00532 b) 0.0879 c) 0.10352

1.7.5 a) 0.3349 b) 0.20094 c) 0.0087

1.7.6 a) $(1-p)^6$, kde $p = p_2 + p_4 + p_6$ b) $15p^2(1-p)^4$
c) $\sum_{i=4}^{6} \binom{6}{i} p^i (1-p)^{6-i}$

1.7.7 a) $4.52\,10^{-5}$ b) 0.002263 c) 0.7866

1.7.8 a) 0.5761 b) 0.007269

1.7.9 a) 0.2821 b) 0.3232

1.7.10 a) 0.8183 b) 0.001026 c) 0.99995

1.7.11 a) 0.0339 b) 0.45129

1.7.12 0.63849

1.7.12 0.63849

1.7.13 a) 56/144 b) 28/144 c) 45/144

1.7.14 a) 0.048 b) 0.000864 c) 0.002592
d) 0.004104 e) 0.03204

1.7.15 0.414875

Kapitola 2

2.1.1 $f(k) = \binom{4}{k}\left(\frac{1}{2}\right)^4$, $k = 0, 1, 2, 3, 4$. $f(0) = f(4) = 0.0625$, $f(1) = f(3) = 0.25$, $f(2) = 0.375$

2.1.2 $f(k) = \binom{6}{k}\left(\frac{1}{3}\right)^k\left(\frac{2}{3}\right)^{6-k}$, $k = 0, 1, \ldots, 6$. $P(X < 4) = 0.8999$

2.1.3 $f(k) = \binom{50}{k}(0.02)^k(0.98)^{50-k}$, $k = 0, 1, \ldots, 50$. $P(X \le 3) = 0.9822$, $P(X > 3) = 0.0178$

2.1.4 $P(X = k) = \dfrac{\binom{3}{k}\binom{7}{4-k}}{\binom{10}{4}}$, $k = 0, 1, 2, 3$. Rozdelenie Y je rovnaké ako rozdelenie X

$P(Z = k) = \binom{4}{k}(0.3)^k(0.7)^{4-k}$, $k = 0, 1, 2, 3, 4$.

2.1.5 $f(k) = P(X = k) = \frac{10-k}{45}$, $k = 1, 2, \ldots, 9$. $P(2 \le X < 6) = P(2 \le X \le 5) = \frac{26}{45}$

2.1.6 $f(5) = f(10) = 1/32$, $f(6) = f(9) = 5/32$, $f(7) = f(8) = 10/32$

2.1.7 $f(0) = 0.0036$, $f(1) = 0.0456$, $f(2) = 0.2116$, $f(3) = 0.4256$, $f(4) = 0.3136$

2.1.8 $f(0) = 0.21$, $f(1) = 0.44$, $f(2) = 0.29$, $f(3) = 0.06$

2.1.9 $f(1) = 6/10$, $f(2) = 4/15$, $f(3) = 1/10$, $f(4) = 1/35$, $f(5) = 1/210$

2.1.10 $f(2) = 1/3$, $f(3) = 1/3$, $f(4) = 9/42$, $f(5) = 4/42$, $f(6) = 1/42$

2.1.11 $f_X(3) = 0.1$, $f_X(4) = 0.3$, $f_X(5) = 0.6$ $f_Y(1) = 0.6$, $f_Y(2) = 0.3$, $f_Y(3) = 0.1$

$H(Z)$	6	7	8	9	10	11	12
$f_Z(z)$	0.1	0.1	0.2	0.2	0.2	0.1	0.1

2.1.12 $P(X = k) = \dfrac{\binom{16}{k}\binom{36}{5-k}}{\binom{52}{5}}$, $k = 0, 1, \ldots, 5$.

2.1.13

$H(X)$	-5	-3	-1	1	3	5
$f(x)$	q^5	$5q^4p$	$10q^3p^2$	$10q^2p^3$	$5qp^4$	p^5

2.1.14

$H(X)$	0	10	20	30	40	50	60
$f(x)$	0.0723	0.0868	0.2517	0.1782	0.2517	0.0868	0.0723

2.1.15 $f(0) = 1/32$, $f(1) = 5/32$, $f(2) = 10/32$, $f(3) = 10/32$, $f(4) = 5/32$, $f(5) = 1/32$.

2.1.16 $f(0) = 0.367879$, $f(1) = 0.367882$, $f(2) = 0.183929$, $f(3) = 0.061343$,
$f(4) = 0.015278$, $f(5) = 0.003125$, $f(6) = 0.000463$, $f(7) = 9.9\ 10^{-5}$, $f(9) \approx 3\ 10^{-6}$

2.1.17 $P(X = k) = \left(\frac{1}{2}\right)^k$, $k = 1, 2, 3, \ldots$

2.1.18 $P(X = 1) = 0.6$, $P(X = k) = 0.4^{k-1}\,0.6$, $k = 2, 3, 4, \ldots$

2.1.19 $P(X = 1) = \frac{1}{3}$, $P(X = k) = \left(\frac{2}{3}\right)^{k-1}\left(\frac{1}{3}\right)$, $k = 2, 3, 4, \ldots$

2.1.20 $P(X = k) = (k-1)\frac{1}{36}\left(\frac{5}{6}\right)^{k-2}$, $k = 2, 3, 4, \ldots$

2.2.1 V tejto úlohe k predstavuje konkrétnu hodnotu: a) $k = 2$, b) $k = 0.5$

2.2.2 Teraz a je parameter. Hodnotou parametra a môže byť ľubovoľné kladné číslo.

2.2.3 a) $a = 1$ (a nie je parameter) b) 0.75
c) $m = 0$ d) $c = 0.5528$

2.2.4 a) $a = 2/3$ (a nie je parameter) b) 0.5833
c) $c = 1.85$

2.2.5 a) $a = 3$ (a nie je parameter) b) 0.125 c) $c = 0.9283$

2.2.6 a) $a = 6$ (a nie je parameter) b) 0.6875 c) $m = 0.5$

2.2.7 a) $a = 1/6$ (a nie je parameter) b) 0.625

2.2.8 a) $a = 1/\pi$ (a nie je parameter) b) 0.5 c) $m = 0$

2.2.9 a) $a = 1$ (a nie je parameter) b) 0.3935

2.2.10 a) a je parameter, hodnotou a môže byť ľubovoľné kladné číslo
b) 0.75
c) $m = \frac{2-\sqrt{2}}{2}a$

2.2.11 a) a je parameter, hodnotou a môže byť ľubovoľné číslo z intervalu $\langle -1, 1 \rangle$
b) $\frac{2-a}{4}$

2.2.12 a) a je parameter, hodnotou parametra môže byť ľubovoľné kladné číslo
b) 0.75
c) $c = 0.5528a$

2.2.13 b) $m = \frac{\ln 2}{\lambda}$

2.3.1 $F(2)=0.3,\ \ F(3)=0.7,\ \ F(8)=0.9.$

2.3.2 pre úlohu 2.2.2 je distribučná funkcia F daná vzťahmi:

$F(x)=0,\ (x\leq 0),\ \ F(x)=x^{a}\ \ (0<x\leq 1),\ \ F(x)=1\ \ (x>1).$

pre úlohu 2.2.3 je funkcia F daná vzťahmi:

$F(x)=0,\ x\leq -1,\ F(x)=\frac{x^2}{2}+x+\frac{1}{2}\ \ (-1\leq x\leq 0),$

$F(x)=\frac{-x^2}{2}+x+\frac{1}{2}\ \ (0\leq x\leq 1)$, nakoniec, $F(x)=1$, pre $x>1$.

pre úlohu 2.2.4 distribučná funkcia F je daná vzťahmi:

$F(x)=0\ \ (x\leq 0),\ \ F(x)=\frac{x^2}{3}\ \ (0<x\leq 1),\ \ F(x)=\frac{2x}{3}-\frac{1}{3}\ \ (1\leq x\leq 2),\ F(x)=1,\ x>2.$

pre úlohu 2.2.9 distribučná funkcia F je daná vzťahmi:

$F(x)=0$, pre $x\leq 0$, $F(x)=1-e^{\frac{-x^2}{2}}$, pre $x>0$.

2.3.4 a) $a=1$, b) $f(x)=\frac{e^{-x}}{(1+e^{-x})^2},\ x\in\mathrm{R}$, c) $x_{0.5}=0$.

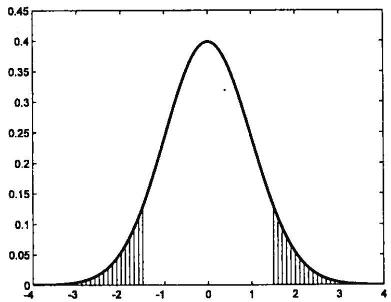
2.4.1 Pre $X\sim\Delta(-2,0,2)$, $F(x)=0$, pre $x<-2$, $F(x)=(x^2+4x+4)/8$, pre $-2\leq x<0$,

$F(x)=(-x^2+4x+4)/8$, pre $0\leq x<2$, $F(x)=1$, pre $x>2$.

2.4.2 $F(x)=0$, pre $x\leq a$, $F(x)=\frac{x-a}{b-a}$, pre $a<x\leq b$, nakoniec, $F(x)=1$, pre $x>b$.

2.4.3 $x_{0.9}=1.1513$

2.4.4 Vyznačená plocha vľavo má obsah rovný $F(x)$. Vyznačená plocha napravo, má plošný obsah, ktorý sa rovná rozdielu celkového, jednotkového obsahu pod Gaussovou krivkou a hodnoty $F(x)$. Preto pre $x<0$ máme $F_N(x)=1-F_N(-x)$.



Plošný obsah plochy vyznačenej vľavo sa rovná $F_N(-1.5)$. Plošný obsah plochy vyznačenej vpravo sa rovná: Celkový plošný obsah pod Gaussovou krivkou, mínus $F_N(1.5)$. Vďaka symetrii sú plošné obsahy rovnaké. Preto $F_N(-1.5)=1-F_N(1.5)$.

2.4.5 a) 0.81859 b) 0.47725 c) 0.69146 d) 0.81859

2.5.1

y_i	-5	-1	3	5	7
$g(y_i)$	0.1	0.2	0.4	0.2	0.1

z_i	0	1	4	16
$g(z_i)$	0.4	0.2	0.3	0.1

Veličina U je konštanta 0 (degenerovaná náhodná veličina). $P(V=-1)=0.2$, $P(V=1)=0.8$

2.5.2 $H(Y)=\{0,1,4\}$, $P(Y=0)=0.2$, $P(Y=1)=0.4$, $P(Y=4)=0.4$, $Z\sim R\{-8,-1,0,1,8\}$

2.5.3 $H(Y)=\{0,1,4,..,(n-1)^2\}$, $P(Y=0)=\frac{1}{2n-1}$, $P(Y=k)=\frac{2}{2n-1}$, $k=1,4,..,(n-1)^2$

$Z\sim R\{(1-n)^3,\ (2-n)^3,\ \ldots,\ -8,\ -1,\ 0,\ 1,\ 8,\ \ldots,\ (n-2)^3,\ (n-1)^3\}$.

2.5.4 $H(Y)=\{5,6,7,\ldots\}$, $P(Y=5)=1-(1-p)^6$, $P(Y=k)=p(1-p)^k$, pre $k=6,7,8,\ldots$

$H(Z)=\{1,\frac{1}{2},\frac{1}{3},\frac{1}{4},\ldots\}$, $P(Z=\frac{1}{k+1})=p(1-p)^k$, $k=0,1,2,\ldots$

2.5.5 $H(Y)=\{0,1,4,9,\ldots\}$, $P(Y=k^2)=e^{-\lambda}\frac{\lambda^k}{k!}$, $k=0,1,2,\ldots$

$H(Z)=\{-1,0,1\}$,

$P(Z=-1)=e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{4k-1}}{(4k-1)!}$, $P(Z=0)=e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!}$, $P(Z=1)=e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^{4k+1}}{(4k+1)!}$

2.5.6 $X\sim N(0,1)\Rightarrow 2X\sim N(0,4)$ a $Y=3+2X\sim N(3,4)$, $-3X\sim N(0,9)$, $Z\sim N(2,9)$.

2.5.7 $F_N(2)-F_N(-1)=0.81859$, $Y\sim N(0,1)$

2.5.8 $a=-0.75$, $b=0.25$

2.5.9 $F_N(1)+F_N(1.5)-1=0.77453$, $Y\sim N(1,1)$.

2.5.10 $Y\sim N(13,36)$, $Z\sim N(3,1)$. Uvedomme si, že $-2X\sim N(6,36)$, $\frac{X}{3}\sim N(-1,1)$.

2.5.11 $F_Y(y)=0$, ak $y\leq 0$, $F_Y(y)=\frac{\sqrt{y}}{4}$, ak $y\in(0,16)$, $F_Y(y)=1$, ak $y\geq 16$

$F_Z(z)=0$, ak $z\leq 0$, $F_Z(z)=\frac{z^2}{4}$, ak $z\in(0,2)$, $F_Z(z)=1$, ak $z\geq 2$.

2.5.12 $Y\sim R(a,a+c)$, $F_Z(z)=0$, ak $z\leq 0$; $F_Z(z)=\frac{\sqrt{z}}{c}$, ak $z\in(0,c^2)$; $F_Z(z)=1$, ak $z\geq c^2$.

2.5.13 $Y\sim\Delta(-2,0,2)$, $Z\sim\Delta(0,1,2)$,

$F_U(u)=0$, ak $u\leq 0$; $F_U(u)=2\sqrt{u}-u$, ak $0<u\leq 1$; $F_U(u)=1$, ak $u>1$.

2.5.14 $Y\sim R(0,1)$

2.5.15 $Y\sim\Delta(-a,0,a)$,

$F_Z(z)=0$, ak $z\leq 0$; $F_Z(z)=\frac{1}{a}\left(2\sqrt{z}-\frac{z}{a}\right)$, ak $0<z\leq a^2$; $F_Z(z)=1$, ak $z>a^2$.

2.5.16 Zrejme $H(Y)=(0,1)$. Pre $y\in(0,1)$ máme $F_Y(y)=P(\sqrt{X}<y)=P(X<y^2)=y^2$, a to je distribučná funkcia rozdelenia $\Delta(0,1,1)$.

2.5.17 $Y = cX \sim \text{Exp}(\frac{\lambda}{c})$,

$F_Z(z) = 0$, ak $z \le 0$; $F_Z(z) = 1 - e^{-\lambda z^2}$, ak $z > 0$.

$F_U(u) = 0$, ak $u \le 1$; $F_U(u) = 1 - e^{-\lambda(u-1)}$, ak $u > 1$.

$F_V(v) = 0$, ak $v \le 0$; $F_V(v) = 1 - e^{-\lambda\sqrt{v}}$, ak $v > 0$.

2.5.18 $F_Y(y) = 0$, ak $y \le 1$; $F_Y(y) = \frac{(y-1)^2}{2} + \frac{y-1}{2}$, ak $1 < y \le 2$; $F_Y(y) = 1$, ak $y > 2$.

$F_Z(z) = 0$, ak $z \le 0$; $F_Z(z) = \frac{z^2}{8} + \frac{z}{4}$, ak $0 < z \le 2$; $F_Z(z) = 1$, ak $z > 2$.

$F_U(u) = 0$, ak $u \le 0$; $F_U(u) = \frac{u}{2} + \frac{\sqrt{u}}{2}$, ak $0 < u \le 1$; $F_U(u) = 1$, ak $u > 1$.

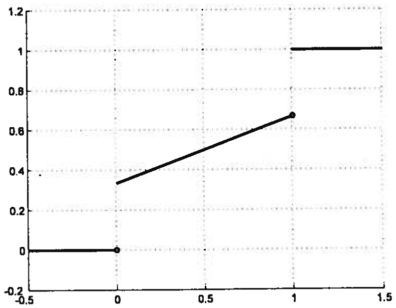
2.5.19 Pre $y \in \langle a,\ b)$ máme: $F_Y(y) = P(Y < y) = P(F^{-1}(U) < y) = P(U < F(y)) = F(y)$.

2.5.20 $P(Y = 0) = P(0 \le X \le 1) = \frac{1}{3}$, $P(Y = 1) = P(2 \le X \le 3) = \frac{1}{3}$,

$P(0 < Y < 1) = P(1 < X < 2) = \frac{1}{3}$.

Veličina Y nie je diskrétna, pretože $H(Y) = \langle 0,\ 1\rangle$, čo znamená, že Y má nespočítateľne veľa hodnôt. Avšak Y nie je ani spojitá s hustotou, pretože v takom prípade by muselo platiť $P(Y = c) = 0$, pre každé $c \in R$ (a to, ako vidíme, neplatí pre dva body – pre $c = 0$ a pre $c = 1$).

Samozrejme, Y je náhodná veličina. Hovoríme, že je *zmiešaného* typu. Na intervale (0, 1) je spojitého typu, ale v bodoch 0, resp. 1 sa chová ako diskrétna veličina (pretože tieto body nadobúda s kladnou pravdepodobnosťou).



Graf distribučnej funkcie veličiny Y na intervale $\langle -0.5, 1.5\rangle$.

Kapitola 3

3.1.1 $a = 0.25$, $b = 0.10$,

$P(X < Y) = 0.30$, $P(X = Y) = 0.30$,

$P(X > Y) = 0.40$, $P(X + Y > 0) = 0.40$.

3.1.2 Napr.

	1	4	9
1	0	$\frac{1}{9}$	$\frac{2}{9}$
2	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
3	$\frac{2}{9}$	$\frac{1}{9}$	0

alebo

	1	4	9
1	0	$\frac{1}{6}$	$\frac{1}{6}$
2	$\frac{1}{6}$	0	$\frac{1}{6}$
3	$\frac{1}{6}$	$\frac{1}{6}$	0

3.1.3 Napr.

	0	1	2	4
0	.2	.1	0	0
1	0	.1	.2	.1
2	0	0	.2	.1

alebo

	0	1	2	4
0	.1	.1	.1	0
1	.1	.1	.1	.1
2	0	0	.2	.1

3.1.4 V tabuľke (pre lepšiu prehľadnosť) namiesto nuly nechávame prázdnu bunku

	2	3	4	5	6	7	8	9	10	11	12
1	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$					
2			$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$				
3					$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$			
4							$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$		
5									$\frac{1}{36}$	$\frac{2}{36}$	
6											$\frac{1}{36}$

3.1.5 Opäť – pre lepšiu prehľadnosť – namiesto núl nechávame prázdne bunky

	2	3	4	5	6	7	8	9	10	11	12
0	$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$
1		$\frac{2}{36}$		$\frac{2}{36}$		$\frac{2}{36}$		$\frac{2}{36}$		$\frac{2}{36}$	
2			$\frac{2}{36}$		$\frac{2}{36}$		$\frac{2}{36}$		$\frac{2}{36}$		
3				$\frac{2}{36}$		$\frac{2}{36}$		$\frac{2}{36}$			
4					$\frac{2}{36}$		$\frac{2}{36}$				
5						$\frac{2}{36}$					

3.1.6 Veličiny U, V majú to isté rozdelenie, ktoré určuje tabuľka:

x	0	1	2
$f(x)$	0.25	0.50	0.25

Rozdelenie (U, V) je dané tabuľkou:

$U \backslash V$	0	1	2
0	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{1}{16}$
1	$\frac{2}{16}$	$\frac{4}{16}$	$\frac{2}{16}$
2	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

3.1.7 Rozdelenie vektora (U, V) určuje tabuľka (rozdelenia zložiek získame na okrajoch) Teraz – na rozdiel od 3.1.6 – sú U, V závislé (charakter tabuľky je celkom iný).

$U \backslash V$	0	1	2	3
0	$\frac{1}{16}$	$\frac{1}{16}$	0	0
1	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	0
2	0	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{1}{16}$
3	0	0	$\frac{1}{16}$	$\frac{1}{16}$

3.1.8 Rozdelenie vektora (X, Y) určuje tabuľka

$X \backslash Y$	0	1	2	3
0	0	0	0	$\frac{1}{120}$
1	0	0	$\frac{21}{120}$	0
2	0	$\frac{63}{120}$	0	0
3	$\frac{35}{120}$	0	0	0

3.1.9 a) Rozdelenie zložiek je rovnaké a je dané funkciou

$$f(x) = \begin{cases} x + 0.5, & 0 < x < 1 \\ 0, & inde \end{cases}$$

b) 0.25
c) 0.50

3.1.10 a) $k = 1/6$
b) Rozdelenia zložiek sú dané hustotami

$$f_1(x) = \begin{cases} \frac{8x-6}{6}, & 1 < x < 2 \\ 0, & inde \end{cases} \qquad f_2(x) = \begin{cases} \frac{6-y}{6}, & 2 < x < 4 \\ 0, & inde \end{cases}$$

c) 0.125
d) $\frac{1}{3}$

3.2.1 X, Y sú závislé, pretože napr. $P(X = 1, Y = 0) \neq P(X = 1)P(Y = 0)$.

3.2.2 a) Úlohe vyhovuje káždá dvojica nezáporných a, b, pre ktoré platí: $a + b = 0.24$
b) Teraz úlohe vyhovuje jediná dvojica: $a = 0.12$, $b = 0.12$

3.2.4 a) $f(x, y) = \frac{1}{12}$, pre $(x, y) \in \{0, 1, 2, 3\} \times \{1, 2, 3\}$, $f(x, y) = 0$, pre ostatné argumenty.
f) $f(i, j) = p_1 p_2 (1 - p_1)^{i-1}(1 - p_2)^{j-1}$, $i, j \in \{0, 1, 2, \ldots, n, \ldots\}$, inak $f(x, y) = 0$.

3.2.6 Veličiny X, Y sú závislé. Obe veličiny majú rozdelenie R(-1, 1).

3.2.7 a) $f(x, y) = \frac{1}{4}$, pre $(x, y) \in (-1, 1) \times (-1, 1)$.

b) $f(x, y) = \lambda^2 e^{-\lambda(x+y)}$, $x > 0, y > 0$, $f(x, y) = 0$, inde.

c) $f(x, y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} = \frac{1}{2\pi} e^{-\frac{(x+y)^2}{2}}$, $x, y \in R$

d) $f(x, y) = \frac{1}{\sqrt{2\pi}\, 2} e^{-\frac{(x-1)^2}{8}} \cdot \frac{1}{\sqrt{2\pi}\, 3} e^{-\frac{(y+2)^2}{18}} = \frac{1}{2\pi\, 6} e^{-\frac{(x-1)^2}{8} - \frac{(y+2)^2}{18}}$, $x, y \in R$

3.2.8 a) $p = (0.84134)^2\, 0.97725 = 0.69175$ b) $p = (0.9545)^3 = 0.8696$

3.3.1 Uveďme, napríklad, rozdelenie veličiny $Z = |X - Y|$. Zrejme $H(Z) = \{0, 1, 2\}$ a pre hodnoty pravdepodobnostnej funkcie platí: $g(0) = 6/15$, $g(1) = 6/15$, $g(2) = 3/15$.

3.3.2 a) Rozdelenie Z určuje tabuľka

z	0	1	2	3	4
$g(z)$	$\frac{4}{81}$	$\frac{20}{81}$	$\frac{33}{81}$	$\frac{20}{81}$	$\frac{4}{81}$

b) $Z \sim Bi(2 + 3, \frac{1}{3}) = Bi(5, \frac{1}{3})$.

3.3.3 a) $Z \sim Bi(3, \frac{1}{2})$
b) Rozdelenie Z určuje tabuľka

z	0	1	2	3
$g(z)$	$\frac{2}{20}$	$\frac{7}{20}$	$\frac{8}{20}$	$\frac{3}{20}$

3.3.4 Ak Z je čas strávený na ceste, tak $Z = Z_1 + Z_2 + Z_3 \sim N(60, 36)$, a preto

a) $P(Z > 70) = 1 - F_N(\frac{70-60}{6}) = 1 - F_N(1.67) = 1 - 0.9525 = 0.0475 \approx 0.05$.

b) $W = X + Y + Z \sim N(10 + 20 + 60,\ 9 + 16 + 36) = N(90, 61)$,

$P(W < 100) = F_N(\frac{100-90}{7.81}) = F_N(1.28) = 0.8997 \approx 0.90$.

3.3.5 $X + Y + Z \sim Erl(3, \lambda)$, $P(X + Y + Z > 300) = 0.4232$.

Pomôcka: neurčitý integrál funkcie $x^2e^{-\lambda x}$ sa rovná $-(\frac{x^2}{\lambda} + \frac{2x}{\lambda^2} + \frac{2}{\lambda^3})e^{-\lambda x}$.

Kapitola 4

4.1.2 Veličiny v úlohách 2.1.1 až 2.1.16 majú konečne veľa hodnôt a ich rozdelenia sú uvedené vo výsledkoch čl.2.1. Výpočet E(X), *var*(X) nepredstavuje problém.

Veličiny z úloh 2.1.17 až 2.1.20 majú rozdelenie, ktoré je modifikáciou rozdelenia G(p) a je treba vedieť niečo z teórie mocninových radov.
Pre zaujímavosť, E(X) = 2, z úlohy 2.1.17, kým E(X) = 1.6667 pre X z úlohy 2.1.18.

4.1.6	E(X)	*var*(X)	*med*(X)	*mkr*(X)
2.2.1a)	0.6667	0.0556	0.7071	0.3660
2.2.1b)	1.3333	0.2222	1.4142	0.7321
2.2.3	0	0.1667	0	0.5858
2.2.4	1.2222	0.2284	1.25	0.759
2.2.5	0.75	0.0375	0.7937	0.2786
2.2.6	0.5	0.05	0.5	0.3473
2.2.7	0	0.2444	0	0.7796
2.2.8	nemá	nemá	0	2
2.2.9	1.2533	0.4292	1.1774	0.9066

4.2.1 $cov(X, Y) = 0.55$, $\rho(X, Y) = 0.6822$.

4.2.2 $E(Y) = 5$, $var(Y) = 36$.

4.2.3 $E(Y) = 8/3$, $var(Y) = 56/9$.

4.2.4	E(X)	*var*(X),	E(Y)	*var*(Y)	*cov*(X, Y)
3.1.4	2.5278	1.97145	7	5.8333	2.9167

4.2.4	E(X)	*var*(X),	E(Y)	*var*(Y)	*cov*(X, Y)
3.1.4	2.5278	1.97145	7	5.8333	2.9167
3.1.5	1.9444	2.0525	7	5.8333	0
3.1.6	1	0.5	1	0.5	0
3.1.7	1.5	0.75	1.5	0.75	0.5
3.1.8	2.1	0.49	0.9	0.49	- 0.49

4.2.5 $a = 3$, $b = 2$, $cov(U, V) = 6\,cov(X, Y)$, $\rho(U, V) = \rho(X, Y)$

4.2.7 a) $E(2X + 3Y) = 5$, $var(2X + 3Y) = 32$
b) $\rho(2X + 3, 4 - X) = -1$
c) $\rho(2X + 3, 1 - Y) = 0.3536$

4.2.8 $var(2X - Y) = 4$.

4.2.9 $cov(X, Y) = -0.37$, $\rho(X, Y) = -0.90175$

4.2.12 a) $E(X + Y) = 16$, $var(X + Y) = 28/3$
b) $E(X - Y) = 8$, $var(X - Y) = 28/3$

4.2.13 a) $E(X) = 7/12$, $var(X) = 11/144$
$E(Y) = 7/12$, $var(Y) = 11/144$, $cov(X, Y) = -1/144$
b) $E(X) = 29/18$, $var(X) = 0.071$,
$E(Y) = 26/9$, $var(Y) = 0.321$, $cov(X, Y) = 1/81$

Kapitola 5

5.1.1 Bez korekcie: 0.57038 ($= F_N(0.64) - F_N(-0.96)$), s korekciou: 0.55303.

5.1.2 Bez korekcie: 0.75800 ($= 2F_N(1.17) - 1$), s korekciou: 0.75507.

5.1.3 Bez korekcie: 0.86324 ($= F_N(1.095)$), resp. 0.86433 ($= F_N(1.10)$).

5.1.4 Ide o určenie $P(S_{150} < 20)$. Avšak $P(S_{150} < 20) \approx F_N(-1.10) = 0.13567$.

5.1.5 $F_N(1.44) - F_N(-1.30) = 0.82827$.

5.2.1 $\frac{S_n - nm}{b\sqrt{n}}$ má tvar $\frac{S_{30} - 30\ 0.5}{\sqrt{\frac{30}{12}}} = \frac{S_{30} - 15}{\sqrt{2.5}}$ a má približne normálne N(0, 1) rozdelenie.

5.2.2 $P(S_{36} \geq 144) = 1 - P(S_{36} < 144) = 1 - F_N(1.76) = 0.03920$,
S_{36} má (približne) rozdelenie N(126, 105).

5.2.3 $P(S_{36} \geq 144) = 1 - P(S_{36} < 144) = 1 - F_N(-1.38) = 0.91621$,
S_{36} má (približne) rozdelenie N(158.4, 109.44).

5.2.4 $P(S_{20} \geq 1000) = 1 - F_N(-2.80) = 0.99744$.

5.2.5 S_n má približne rozdelenie $N(n, 2n)$. $E(X_i^2) = 1$ a $var(X_i^2) = 2$, pretože $E(X_i^4) = 3$.

Kapitola 6

6.6.1 Časť a) je zrejmá. Pri presnom argumentovaní sa opierame sa o čl. 1.7, resp. čl. 3.2.

Časť b) je náročnejšia. Napr. závislosť X_1, X_2, X_3, X_4, X_5 môžeme zdôvodniť tak, že ukážeme, že závislé sú už X_1 a X_2 (to nie je problém, veď tabuľka pravdepodobnostnej funkcie spoločného rozdelenia zrejme obsahuje nuly).

To, že všetky X_i majú rozdelenie R{1, 2, ..., 10}, sa ukáže štandardnými postupmi tak, ako sme to robili pri podmieňovaní (viď čl. 1.5).

6.6.2 Ak $X \sim R\{1, 2, .., N\}$, tak $E(X) = \sum_{i=1}^{N} x_i f(x_i) = \sum_{i=1}^{N} i\,\frac{1}{N} = \frac{1}{N}\sum_{i=1}^{N} i = \frac{1}{N}\frac{N(N+1)}{2} = \frac{N+1}{2}$

Strednú hodnotu E(X) odhadujeme štatistikou $\bar{X}$, preto $\frac{\widehat{N+1}}{2} = \bar{X}$, odkiaľ $\hat{N} = 2\bar{X} - 1$.

Ak realizáciou výberu sú čísla 11, 7, 15, 8, 4, tak $\bar{x} = 9$ a pre odhad máme: $\hat{N} = 17$.

6.6.4 a) $\bar{x} = 5.027$, $s = 2.702$

b) $\bar{x} = 4.993$, $s = 2.644$, teda tieto aproximácie sú presnejšie, ako tie z časti a).

6.6.5 a) $\bar{x} = 115$, $s = 9.8184$

b) $G_n(110) = 3/11$, $G_n(111) = 4/11$, $G_n(121) = 8/11$, $G_n(125) = 9/11$.

c) $F_n(110) = 7/22$, $F_n(111) = 8/22$, $F_n(121) = 17/22$, $F_n(125) = 18/22$.

d) $\tilde{x}_{0.5} = 113$, $\tilde{x}_{0.25} = 109.25$, $\tilde{x}_{0.75} = 120.25$, výberové $mkr = \tilde{x}_{0.75} - \tilde{x}_{0.25} = 11$.

Kapitola 7

7.3.1 Pomôcka: Ak $X_i \sim N(\mu, \sigma^2)$, tak, samozrejme, $E[(X_i - \mu)^2] = var(X_i) = \sigma^2$.

7.3.2 $E(g(X)) = E(h(X)) = \theta$, ale $h(X)$ má menšiu varianciu, a preto je lepším odhadom θ,

$var[g(X)] = \frac{25}{18}\frac{\theta^2}{12} > var[h(X)] = \frac{\theta^2}{12}$.

7.3.3 Lepším odhadom m je $h(X)$, pretože $var[h(X)] = \frac{\sigma^2}{15}$, kým $var[g(X)] = \frac{3\sigma^2}{40}$.

7.3.4 $E([h(X)]^2) = var(h(X)) + (E[h(X)])^2 = var(h(X)) + \theta^2 > \theta^2$, pretože podľa predpokladu $var(h(X)) > 0$ a tiež $E[h(X)] = \theta$. Takto $[h(X)]^2$ je vskutku vychýlený odhad pre θ^2.

7.3.5 $k = 3$ (pomôcka: Ak $X_i \sim R(-\theta, \theta)$, tak $E[X_i^2] = \frac{\theta^2}{3}$.)

Kapitola 8

8.5.1 a) Hladina testu založeného na K sa rovná 0.087, $\beta = 0.250$, teda sila sa rovná 0.75.

b) Podľa CLV, ak H_0 platí, máme $Z \sim N(4, 3.2)$. Ak platí H_1, tak $Z \sim N(8, 4.8)$. Hladina testu sa rovná 0.081, $\beta = 0.248$ (obe hodnoty sú počítané s korekciou). Vidíme, že použitie CLV dalo veľmi dobré aproximácie oboch pravdepodobností.

8.5.2 Ak X je indikátor bielej loptičky, tak $X \sim A(p)$, kde p je pravdepodobnosť vytiahnutia bielej. Test má tvar

$$H_0\colon p = 0.4 \quad \text{proti} \quad H_1\colon p = 0.6.$$

Testovacou štatistikou je súčet $\sum_{i=1}^{5} X_i$ a kritická množina $K = \{x \in R^5 : \sum_{i=1}^{5} x_i \geq 3\}$.

Hladina testu α sa rovná $0.31744 \approx 0.32$. V tomto prípade, vzhľadom na symetriu hypotéz, platí: $\beta = \alpha$. Sila testu sa rovná $1 - 0.32 = 0.68$.

8.5.3 Hladina testu sa rovná $0.0511 \approx 0.05$, $\beta = 0.4529 \approx 0.45$, t.j. sila testu sa rovná 0.55.

8.5.4 Testovacou štatistikou môže byť výberový priemer $\bar{X}$. Kritická hodnota výberového priemeru $\bar{X}$ sa rovná 11.32, pretože

$$K = \{x \in R^{25} : \bar{x} \geq 11.32\}$$

je kritická oblasť veľkosti 0.05.

$\beta = P(\bar{X} < 11.32 \,||\, \mu = 12) = 0.19766$, t. j. $\beta \approx 0.20$ a sila testu sa rovná 0.80. Ak sa výberový priemer realizuje hodnotou 11.23, tak H_0 nezamietame (na hladine 0.05).

8.5.5 $K = \{x \in R^{20} : x_1 + x_2 + \ldots + x_{20} \leq 7\} \cup \{x \in R^{20} : x_1 + x_2 + \ldots + x_{20} \geq 13\}$.

$\alpha = P(X_1 + X_2 + \ldots + X_{20} \in K \,||\, p = 0.5) = 0.2632$,

$\beta(0.6) = P(8 \leq X_1 + X_2 + \ldots + X_{20} \leq 12 \,||\, p = 0.6) = 0.5631$.

8.5.6 Test má tvar $H_0\colon p = \frac{1}{6}$ proti $H_1\colon p \neq \frac{1}{6}$.

Nech X_i je indikátor padnutia 6-tky v i-tom hode. Potom

$$K = \{x \in R^{60} : x_1 + x_2 + \ldots + x_{60} \leq 5\} \cup \{x \in R^{60} : x_1 + x_2 + \ldots + x_{60} \geq 15\},$$

$\alpha = 2(1 - F_N(1.56)) = 0.11876 \approx 0.12$ (uplatnenie CLV s korekciou).

8.5.7 $K = \{x \in R^{25} : |\bar{x} - 10| \geq \frac{4\,u_{0.975}}{\sqrt{25}}\} = \{x \in R^{25} : |\bar{x} - 10| \geq 1.568\} =$

$= \{x \in R^{25} : \bar{x} \leq 8.432\} \cup \{x \in R^{25} : \bar{x} \geq 11.568\}$.

$\beta(12) = P(8.432 \leq \bar{X} \leq 11.568 \,||\, \mu = 12) = 0.2946$, t.j. sila (ak $\mu = 12$) sa rovná $1 - 0.2946 \approx 0.71$.

8.5.8 $K = \{x \in R^{25} : \bar{x} \geq 10 + \frac{4\,u_{0.95}}{\sqrt{25}}\} = \{x \in R^{25} : \bar{x} \geq 11.316\}$.

$\beta(12) = P(\bar{X} \leq 11.316 \,||\, \mu = 12) = F_N(-0.855) = 0.1963$, t.j. sila (ak $\mu = 12$) sa (po zaokrúhlení) rovná 0.80. Porovnajte ju s výsledkom v 8.5.7.

8.5.9 Test má tvar $H_0\colon \mu = 250$ proti $H_1\colon \mu \neq 250$

kritickou oblasťou je podľa 8.3.10 (a):

$K = \{x \in R^{10} : |\bar{x} - 250| \geq \frac{s\,t_{1-\alpha/2}(n-1)}{\sqrt{n}}\} = \{x \in R^{10} : |\bar{x} - 250| \geq \frac{3.0288\ 2.2622}{\sqrt{10}}\} =$

$= \{x \in R^{10} : |\bar{x} - 250| \geq 2.167\}$. Keďže $\bar{x} = 251.54$, tak $x \notin K$, a preto H_0 nezamietame.

8.5.10 Pre test (1) $\quad H_0: \mu \geq 1500$ proti $H_1: \mu < 1500$

je podľa vety 8.3.10 (c), kritickou oblasťou $K = \{x \in R^{12}: \bar{x} \leq \mu_0 - \frac{s\, t_{1-\alpha}(11)}{\sqrt{12}}\} =$

$= \{x \in R^{12}: \bar{x} \leq 1500 - \frac{83.37\ 1.7959}{\sqrt{12}}\} = \{x \in R^{12}: \bar{x} \leq 1456.78\}$.

Pretože $\bar{x} = 1469.3$, $x \notin K$, H_0 nezamietame. Prijímame hypotézu, že životnosť je aspoň 1500 hodín. Nevýhodou takejto formulácie testu je, že nevieme, s akou pravdepodobnosťou prijímame neplatnú H_0.

Pre test (2) $\quad H_0: \mu \leq 1500$ proti $H_1: \mu > 1500$

je podľa vety 8.3.10 (b) kritickou oblasťou $K = \{x \in R^{12}: \bar{x} \geq \mu_0 + \frac{s\, t_{1-\alpha}(11)}{\sqrt{12}}\} =$

$= \{x \in R^{12}: \bar{x} \geq 1500 + \frac{83.37\ 1.7959}{\sqrt{12}}\} = \{x \in R^{12}: \bar{x} \geq 1543.22\}$ a keďže $\bar{x} = 1469.3$,

$x \notin K$, H_0 nezamietame. Prijímame teda hypotézu, že životnosť je najviac 1500 hodín. Priali sme si zamietnuť H_0, ale, žiaľ, naše dáta to neumožňujú.

Vziať závery testov (1), (2) do úvahy a prehlásiť, že prijímame hypotézu $\mu = 1500$, vyzerá na pohľad logické, ale nie je to korektné. Teda nie v zmysle štatistického testovania, v ktorom hovoríme o pravdepodobnosti možných omylov – pozri poznámku 8.4.6. Namiesto takého prehlásenia je korektné previesť test $H_0: \mu = 1500$ proti $H_1: \mu \neq 1500$.

8.5.11 (a) $K = \{(x, y) \in R^{m+n}: |\bar{x} - \bar{y}| \geq \frac{\sigma u_{1-\alpha/2}}{\sqrt{\frac{mn}{m+n}}}\}$

(b) $K = \{(x, y) \in R^{m+n}: \bar{x} \geq \bar{y} + \frac{\sigma u_{1-\alpha}}{\sqrt{\frac{mn}{m+n}}}\}$

8.5.12 $E(S_p^2) = \frac{1}{m+n-2}((m-1)E(S_X^2) + (n-1)E(S_Y^2)) = \frac{1}{m+n-2}((m-1)\sigma^2 + (n-1)\sigma^2) = \sigma^2$.

8.5.13 Pre test $\quad H_0: \mu_1 \leq \mu_2$ proti $H_1: \mu_1 > \mu_2$

je oblasť $\quad K = \{(x, y) \in R^{25}: \bar{x} - \bar{y} \geq \frac{s_p\, t_{0.95}(df)}{\sqrt{\frac{154}{14+11}}}\}$

kritickou oblasťou veľkosti (približne) 0.05. Chceme ilustrovať, že ak variancia súborov je rovnaká (v našom prípade rovná 0.25), počet stupňov voľnosti *df* podľa Satterthwaitovho vzťahu sa rovná (približne) $m + n - 2$ (= 23).

Z dát (x_i), (y_i) máme: $s_x = 0.503$, $s_y = 0.499$, a preto

$$\frac{[(S_X^2/m) + (S_Y^2/n)]^2}{\frac{(S_X^2/m)^2}{m-1} + \frac{(S_Y^2/n)^2}{n-1}} = \frac{\left[\frac{0.503^2}{14} + \frac{0.499^2}{11}\right]^2}{\frac{(0.253/14)^2}{13} + \frac{(0.249/11)^2}{10}} = \frac{(0.01807+0.02264)^2}{0.00002512+0.00005124} = \frac{0.0016573}{0.00007636} = 21.70,$$

a po zaokrúhlení $df = 22$. Pre realizovanie testu potrebujeme odhad smerodajnej odchýlky, ktorá je smerodajnou odchýlkou oboch súborov.

$$s_p = \sqrt{\frac{1}{23}(13s_x^2 + 10s_y^2)} = \sqrt{\frac{13\ 0.253 + 10\ 0.249}{23}} = \sqrt{0.25126} = 0.5013.$$

Výraz $\frac{s_p\, t_{0.95}(df)}{\sqrt{\frac{154}{14+11}}} = \frac{0.5013\ 1.7171}{2.4819} = 0.3468$, keďže $\bar{x} - \bar{y} = 5.981 - 5.404 = 0.577$, to znamená,

že (x, y) patrí do K, a preto H_0 zamietame (čo sme, samozrejme, očakávali).

8.5.14 Pre test $\quad H_0: \mu_1 \leq \mu_2$ proti $H_1: \mu_1 > \mu_2$

je kritickou oblasťou veľkosti (približne) 0.05 množina

$$K = \{(x, y) \in R^{28}: \bar{x} - \bar{y} \geq \frac{s_p\, t_{0.95}(df)}{\sqrt{\frac{196}{14+14}}}\}$$

Z realizácií (x_i), (y_i) máme $\bar{x} = 59.8714$, $s_x = 0.7311$, $\bar{y} = 55.7143$, $s_y = 1.0669$, preto

$$s_p = \sqrt{\frac{1}{26}(13s_x^2 + 13s_y^2)} = \sqrt{\frac{13\ 0.5345 + 13\ 1.1383}{26}} = \sqrt{0.8364} = 0.9145.$$

$$\frac{[(S_X^2/m) + (S_Y^2/n)]^2}{\frac{(S_X^2/m)^2}{m-1} + \frac{(S_Y^2/n)^2}{n-1}} = \frac{\left[\frac{0.5345}{14} + \frac{1.1383}{14}\right]^2}{\frac{(0.5345/14)^2}{13} + \frac{(1.1383/14)^2}{13}} = \frac{(0.0382+0.0813)^2}{0.000112+0.0005085} = 23.014, \text{ t.j } df = 23.$$

Výraz $\frac{s_p\, t_{0.95}(df)}{\sqrt{\frac{196}{28}}} = \frac{0.9145\ 1.7139}{2.64575} = 0.5924$, keďže $\bar{x} - \bar{y} = 59.8714 - 57.7143 = 2.1571$, zrejme

$(x, y) \in K$, a preto na hladine 0.05 H_0 zamietame (prijímame teda hypotézu, že pevnosť vlákna sa znížila).

8.5.15 V tejto situácii ide o párový *t*-test, teda nie o dvojvýberový test. Veličina *X* predstavuje čas, ktorý potrebuje prvý kompilátor na skompilovanie náhodne vybratého programu a *Y* je čas, ktorý potrebuje na skompilovanie *toho istého* programu druhý kompilátor. Zrejme veličiny *X*, *Y* sú závislé, pretože sa vzťahujú na ten istý objekt (na ten istý program).

Chceme porovnať E(X) a E(Y). Položíme $Z = X - Y$ a budeme testovať strednú hodnotu E(Z). Ide o jednovýberový *t*-test o strednej hodnote μ veličiny *Z*, o ktorej vieme, že platí: $Z \sim N(\mu, \sigma^2)$, $\mu = E(Z) = E(X) - E(Y)$. Vzhľadom na okolnosti, test má tvar

$$H_0: \mu \leq 0 \quad \text{proti} \quad H_1: \mu > 0$$

v situácii, keď parameter σ nepoznáme – preto postupujeme podľa vety 8.3.10 (b), $\mu_0 = 0$.

Radi by sme H_0 zamietli. To by podporilo tvrdenie, že druhý kompilátor pracuje rýchlejšie (zrejme $\mu = E(X) - E(Y) > 0$ znamená, že $E(X) > E(Y)$). Našimi dátami sú z_i, $z_i = x_i - y_i$:

-0.04 0.21 0.24 0.14 0.16 0.22 0.09 -0.06 -0.06 0.14

Kritickou oblasťou je $K = \{z \in R^{10}: \bar{z} \geq \frac{s\, t_{1-\alpha}(n-1)}{\sqrt{n}}\} = \{z \in R^{10}: \bar{z} \geq \frac{s\, t_{0.95}(9)}{\sqrt{10}}\}$, kde *s* je odhad smerodajnej odchýlky σ, veličiny Z.

Keďže $s = \sqrt{\frac{1}{9}\sum_{i=1}^{10}(z_i - \bar{z})^2} = 0.1172$ a $t_{0.95}(9) = 1.8331$, máme

$$\frac{s\,t_{0.95}(9)}{\sqrt{10}} = \frac{0.1172\ 1.8331}{3.1623} = 0.0679.$$

Pretože $\bar{z} = 0.104$ zisťujeme, že $z \in K$, a to znamená, že na hladine 0.05 hypotézu H_0 zamietame. Na hladine 0.05 teda prijímame alternatívnu hypotézu, že druhý kompilátor kompiluje rýchlejšie.

LITERATÚRA

[1] Anděl, J.: Statistické metody. MatFyzPress, Praha, 1998.

[2] Bhattacharyya, G. K., Johnson, R. A.: Statistical Concepts and Methods. John Wiley & Sons, New York, 1977.

[3] Čistjakov, V. P.: Kurs teorii verojatnostej. Nauka, Moskva, 1987.

[4] Dekking, F. M. et al.: A Modern Introduction to Probability and Statistics. Springer-Verlag, London, 2005.

[5] Hogg, R. V., Ledolter, J.: Engineering Statistics. Macmillan Publishing Company, New York, 1987.

[6] Likeš, J., Machek, J.: Počet pravděpodobnosti. SNTL, Praha, 1981.

[7] Likeš, J., Machek, J.: Matematická statistika. SNTL, Praha, 1983.

[8] Rényi, A.: Teorie pravděpodobnosti. Academia, Praha, 1972.

[9] Riečanová, Z. a kol.: Numerické metódy a matematická štatistika. SNTL/Alfa, Bratislava, 1987.

[10] Ross, S. M.: Introduction to Probability Models. Academic Press, London, 1997.

[11] Svešnikov, A. A. a kol.: Sbírka úloh z teorie pravděpodobnosti a matematické statistiky. SNTL, Praha, 1971.

[12] Tutubalin, V. N.: Teorie pravděpodobnosti. SNTL, Praha, 1978.

[13] Volauf, P.: Matematická štatistika (zbierka príkladov). Alfa, Bratislava, 1988.

[14] Zvára, K., Štěpán, J.: Pravděpodobnost a matematická statistika. MatFyzPress, Praha, 2006.

REGISTER

Doc. RNDr. Peter Volauf, PhD.

PRAVDEPODOBNOSŤ A MATEMATICKÁ ŠTATISTIKA

Vydala Slovenská technická univerzita v Bratislave v Nakladateľstve STU, Bratislava, Vazovova 5, v roku 2014.

Edícia vysokoškolských učebníc

Rozsah 203 strán, 55 obrázkov, 30 tabuliek, 12,836 AH, 13,126 VH, 1. vydanie, edičné číslo 5764, tlač Nakladateľstvo STU v Bratislave.

85 – 212 – 2014

ISBN 978-80-227-4144-6