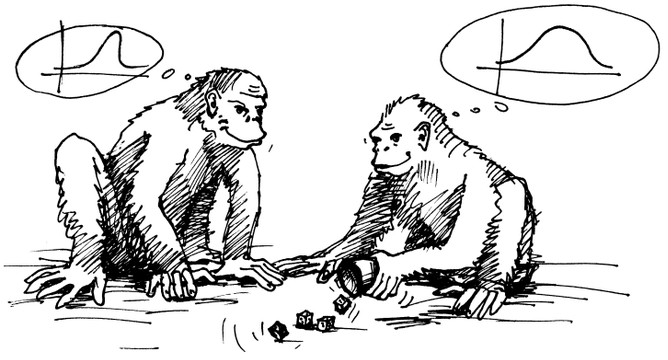
KAPITOLA 9

Statistické a pravděpodobnostní metody

Je statistika částí matematiky?
– když ano, pak matematiky potřebuje moc…!



A. Tečky, čáry, obdélníčky

Získaná data z praxe můžeme zachytit různými způsoby. Uveďme několik základních.

9.1. Zobrazování získaných dat. U 20 matematiků bylo zjištěn počet členů domácnosti, ve které žijí. V tabulce je uvedena četnost, se kterou se dané počty členů domácnosti vyskytly.

Počet členů	1	2	3	4	5	6
Počet domácností	5	5	1	6	2	1

Vytvořte tabulku rozložení četnost. Určete průměr, medián a modus počtu osob v domácnosti. Sestavte sloupcový diagram dat.

Řešení. Do tabulky rozložení četností zapíšeme nejen vlastní četnosti, ale i kumulativní četnosti a pravděpodobnosti, že s jakou má náhodně vybraná domácnost daný počet členů (tzv. relativní četnost). Možný počet členů domácnosti označíme x_i, odpovídající četnost pak n_i, relativní četnost p_i $(= n_i / \sum_{j=1}^{6} n_j = n_i/20)$, kumulativní četnost N_i $(= \sum_{j=1}^{i} x_j)$ a relativní kumulativní četnost

Statistika je, v širším slova smyslu, jakékoliv zpracování číselných nebo jiných dat o nějakém souboru objektů a jejich více či méně přehledná prezentace. V tomto smyslu hovoříme o *popisné statistice*. Jejím předmětem je tedy zpracování a zpřehledňování dat o objektech daného souboru, např. roční příjmy všech občanů zpracovávané z kompletních dat finančních úřadů.

Matematická statistika spočívá ve využití matematických metod pro odvozování závěrů platných pro celý (potenciálně nekonečný) soubor objektů na základě nějakého „malého“ vzorku. Např. zjišťujeme zatížení populace chorobami pomocí dat získaných u několika nahodile vybraných osob, chceme ale interpretovat výsledky ve vztahu k celé populaci.

Podstatou popisné statistiky je odvození jednoduchých (zpravidla) číselných charakteristik o velkých souborech dat, resp. jejich vhodná vizualizace. Podstatou matematické statistiky je pro prezentovaná data zjišťovat, jaké vlastnosti skutečně mají objekty, které jsou daty popisovány, a zároveň, jak věrohodné jsou odvozené výsledky. Zpravidla přitom jde o sběr a zpracování dat o nějakém souboru objektů, jejich následnou analýzu a, konečně, o vyslovení důsledků pozorování pro rozsáhlejší soubor objektů než jsou ty, jejichž data jsme zpracovávali. Ještě jinak řečeno, výsledkem použití matematické statistiky je sdělení o velkém souboru objektů na základě studia malé (zpravidla náhodně vybrané) části z nich, společně s kvalitativním odhadem věrohodnosti výsledného sdělení.

Matematická statistika je opřena hlavně o nástroje teorie pravděpodobnosti, které jsou velice užitečné (a zajímavé) i samy o sobě. Nejvíce úsilí budeme v dalším textu věnovat právě jim.

Celá tato kapitola poskytuje elementární úvod do metod pravděpodobnosti a statistiky, který by měl být dostatečný pro správné chápání běžných statistických informací všude kolem nás. Pro seriozní porozumění práci matematického statistika bude třeba sáhnout po dalších zdrojích.

1. Popisná statistika

Popisná statistika není sama o sobě matematická disciplína, byť používá četné manipulace s čísly a občas i velmi sofistikované metody. Je přitom ale dobrou příležitostí k ilustraci matematického přístupu k budování obecně užitečných nástrojů.

Zároveň by nám měla posloužit jako motivace pro řadu úvah v pravděpodobnosti, protože už budeme tušit, k čemu je později v matematické statice budeme potřebovat.

$F_i (= N_i/20 = \sum_{j=1}^{i} p_j)$:

x_i	n_i	p_i	N_i	F_i
1	5	1/4	5	1/4
2	5	1/4	10	1/2
3	1	1/20	11	11/20
4	6	3/10	17	17/20
5	2	1/10	19	19/20
6	1	1/20	20	1

Snadno již také sestavíme požadované (sloupcové) grafy (relativních, kumulativních) četností:

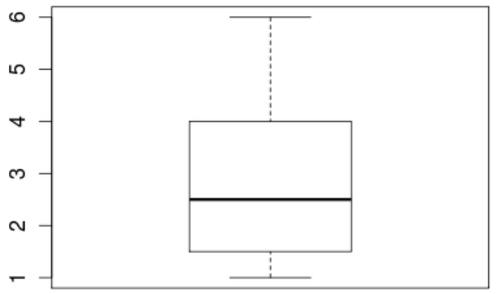
Snadno spočítáme průměr počtu osob v domácnosti:

$$\bar{x} = \frac{5 \cdot 1 + 5 \cdot 2 + 3 \cdot 1 + 6 \cdot 4 + 2 \cdot 5 + 1 \cdot 6}{20} = 2{,}9.$$

Medián je pak průměr desáté a jedenácté hodnoty (seřazených podle velikosti), tedy průměr z čísla 2 a 3, $\tilde{x} = 2, 5$.

Modus je nejčastěji se vyskytující hodnota, tedy $\hat{x} = 4$.

Uvedená data také můžeme zobrazit pomocí *krabicového diagramu*:



Horní a dolní strana „krabice“ odpovídá prvnímu (též dolnímu), resp. třetímu (též hornímu) kvartilu, její výška je tedy rovna kvartilovému rozpětí. Tlustá vodorovná čára mediánu je vedena ve výšce mediánu, dolní a horní vodorovná čára v diagramu odpovídá minimálnímu a maximálnímu prvku výběru, případně hodnotě, která je o 1,5 násobku kvartilového rozpětí nižší (resp. vyšší) než dolní (resp. horní) strana krabice. Případná data mimo toto rozpětí značíme v diagramu kolečky.

Není též problém sestavit histogram daných dat:

9.1. Pravděpodobnost nebo statistika? Ne náhodou se vracíme

k části našich motivačních náznaků z první kapitoly, jak jen se nám podařilo shromáždit dostatek matematických nástrojů jak diskrétní, tak spojité povahy.

Statistikami je totiž dnes zaplaveno kdejaké sdělení, ať už v médiích, politické nebo odborné. Nicméně porozumět obsahu takového sdělení a pochopit možnosti či oprávněnost využití jednotlivých statistických metod a pojmů si vyžaduje mnoho znalostí z různých oblastí matematiky, kterými jsme dosud procházeli. V tomto odstavci ještě nezačneme s matematickou teorií — ve volném sledu poznámek se jen zamyslíme nad dalšími kroky a cíli.

Vezměme si jako příklad souboru objektů všechny studenty konkrétního základního kurzu. Jako číselné údaje pak můžeme např. zkoumat

- „průměrný počet bodů“ dosažený při hodnocení tohoto předmětu v minulém semestru a „rozptyl“ dosažených hodnot,
- „průměrné známky“ dosažené u zkoušky z tohoto a z jiných pevně vybraných předmětů a „korelace“ (tj. vzájemnou souvislost) mezi výsledky,
- „korelace“ dat vypovídajících o historii dřívějšího studia u konkrétních studentů,
- „korelace“ neúspěchů ve studiu a počtu pracovních hodin týdně odpracovaných studentem či studentkou mimo fakultu,
- ...

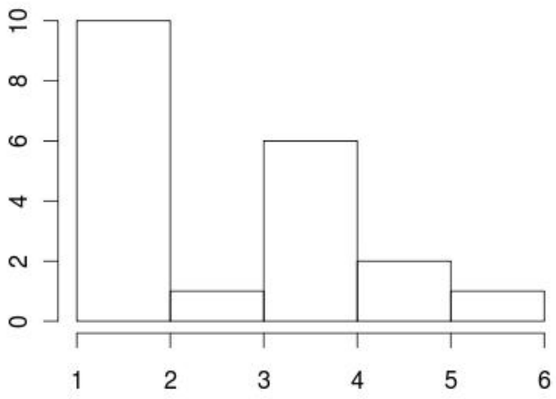
Zastavme se u prvního údaje. Samotný aritmetický průměr bodů nám mnoho neřekne ani o kvalitě přednášky ani o kvalitě přednášejícího ani o samotném hodnocení konkrétních studentů. Možná nás bude více zajímat hodnota, která bude „uprostřed souboru“, tj. počet bodů, pro které je stejně studentů pod ní a nad ní (nebo obdobně první a poslední čtvrtina, desetina apod.). Všem takovým údajům říkáme *statistiky* posuzované veličiny. Takové statistiky budou jistě zajímavé pro samotné studenty a je docela jednoduché je zavést, spočíst i sdělit.

Z obecné zkušenosti nebo jako výsledek teoretických úvah mimo samotnou matematiku víme, že rozumné hodnocení by mělo mít tzv. „normální“ rozdělení. Tento pojem patří do teorie pravděpodobnosti a k jeho zavedení potřebujeme poměrně dost matematiky. Porovnáním výsledku třeba i docela malého náhodného výběru studentů s teoretickým předpokladem můžeme zjistit odhad parametrů takového rozdělení, ale také činit závěry, zda je celé hodnocení postaveno rozumně.

Zároveň lze z číselných hodnot našich statistik pro konkrétní výběr kvalitativně popsat věrohodnost našich závěrů. Stejně tak budeme umět spočíst statistiky, které nebudou odrážet polohy hodnot uvnitř daného statistického souboru ale variabilitu sledovaných hodnot. Tak například když výsledky hodnocení nebudou vykazovat dostatečnou variabilitu, přičemž studenti jistě různé výkony prokazují, jde opět o náznak, že je s předmětem něco v nepořádku. Když působí zjištěná data zcela chaotickým dojmem, pak asi také.

V předchozím odstavci jsme mlčky předpokládali, že považujeme zpracovávaná data za věrohodná. To však v praktickém využití tak nebývá. Naopak samotná data jsou zatížena chybami, zpravidla vznikajícími v důsledku konstrukce experimentu a samotného sběru dat.

V mnoha případech také není známo mnoho o charakteru rozdělení dat. V takových případech je obvyklé používat metody



Všimněme si, že četnosti výskytů jedno a dvoučlenných domácností byly sloučeny do jednoho obdélníčku. Tento postup se používá pro „zpřehlednění dat" – existují (různá a nejednoznačná) pravidla, jak při slučování postupovat. Proto pouze na tento fakt upozorňujeme, aniž bychom uvedli přesný postup (v podstatě je to, jak se to komu líbí). □

9.2. Pro soubor znaků $x = (x_1, x_2, \ldots, x_n)$ vypočtěte průměr a rozptyl centrovaných hodnot $x_i - \bar{x}$ a standardizovaných hodnot $\frac{x_i - \bar{x}}{s_x}$.

Řešení. Průměr centrovaných hodnot zjistíme přímým výpočtem za použití definice aritmetického průměru

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{\bar{x}}{n}\sum_{i=1}^{n} 1 = \bar{x} - \bar{x} = 0.$$

Rozptyl centrovaných hodnot je zřejmě shodný s rozptylem původních hodnot s_x. Pro standardizované hodnoty je průměr zjevně opět roven nule a rozptyl je roven

$$\tfrac{1}{n}\sum_{i=1}^{n}\left(\tfrac{x_i-\bar{x}}{s_x}\right)^2 = \tfrac{1}{s_x^2}\cdot\tfrac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = 1. \qquad \square$$

9.3. Dokažte, že pro rozptyl platí vztah $s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$.

Řešení. Z definice rozptylu a aritmetického průměru

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{2\bar{x}}{n}\sum_{i=1}^{n} x_i + \bar{x}^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2.$$

□

9.4. Byly naměřeny následující hodnoty nějakého znaku

10; 7; 7; 8; 8; 9; 10; 9; 4; 9; 10; 9; 11; 9; 7; 8; 3; 9; 8; 7.

Určete aritmetický průměr, medián, kvartily, rozptyl a příslušný krabicový diagram.

neparametrické statistiky (kterých se jen letmo dotkneme na konci kapitoly).

Velmi zajímavé vývody můžeme formulovat, když porovnáním statistik pro různé veličiny uvedené výše budeme moci dovozovat informace o souvislostech. Pokud např. neexistuje žádná doložitelná souvislost mezi historií předchozího studia a výsledky v dané přednášce, je jedním z možných vysvětlení závěr, že je přednáška prostě špatně vedená.

Shrňme si tedy tyto úvahy takto:

- V popisné statistice máme k dispozici nástroje, které umožňují dobře porozumět struktuře a povaze i velmi rozsáhlých dat;
- v matematice pracujeme s abstraktním matematickým popisem pravděpodobnosti, který je použitelný pro analýzu daných dat, zejména když máme k dispozici teoretický model, kterému mají odpovídat;
- závěry statických šetření na vzorcích konkrétních souborů dat může dát matematická statistika;
- i to, do jaké míry je takový popis adekvátní pro konkrétní výběr dat, je možné vyjádřit pomocí metod matematické statistiky.

Než se do takového složitého programu pustíme, zastavme se u prvního bodu.

9.2. Terminologie. Statistikové zavedli veliké množství názvů a budeme si je muset osvojit. Základním východiskem je *statistický soubor*, což je přesně definovaná množina základních *statistických jednotek*. Ty mohou být dány buď výčtem nebo nějakými pravidly v rámci většího souboru.

Na každé statistické jednotce měříme jeden nebo více *statistických znaků*, přitom ovšem chápeme „měření" velice široce.

Např. souborem mohou být všichni studenti dané univerzity, každý zvlášť je pak *statistickou jednotkou*. O těchto jednotkách pak můžeme schraňovat mnoho znaků – např. všechny číselné hodnoty zjistitelné z informačního systému, jakou mají jednotliví studenti nejraději barvu, co snědli večer před poslední písemkou, atd.

Základním objektem pro zkoumání jednotlivých znaků je pak *soubor hodnot*. Zpravidla jej máme ve formě uspořádaných hodnot. Uspořádání je buď dáno přirozeně (když jsou hodnotami např. reálná čísla) nebo je můžeme zavést pro určitost (třeba když budeme sledovat barvy, tak je můžeme vyjadřovat v RGB standardu a řadit podle tohoto příznaku). Můžeme pracovat i s hodnotami neuspořádanými.

Protože smyslem statistického popisu je srozumitelně a přehledně sdělit něco o celém souboru, budeme jistě chtít umět jednotlivé hodnoty nějak porovnávat a poměřovat. Je tedy podstatné mít k tomu dispozici nějaké *měřítko*. Nejčastěji máme znaky vyjádřeny číselnou hodnotou. Ovšem věcný význam dat může být kvantifikován v různé míře a podle toho rozeznáváme různé *typy měřítek* znaků.

Typy měřítek znaků

Podle toho jakého charakteru jsou hodnoty, hovoříme o typu:

- *nominálním*, kdy mezi hodnotami není žádný vztah, jde pouze o označení jednotlivých kvalitativních jmen, tj. možných hodnot (např. politické strany v ČR nebo přednášející na univerzitě při zkoumání jejich obliby);

Řešení. Označíme-li různé hodnoty znaku a_i a jejich četnosti n_i, pak můžeme soubor dat ze zadání uspořádat do následující tabulky.

a_i	3	4	7	8	9	10	11
n_i	1	1	4	4	6	3	1

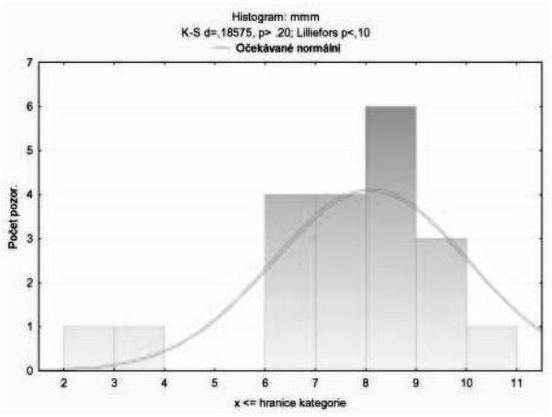
Z definice aritmetického průměru pak máme

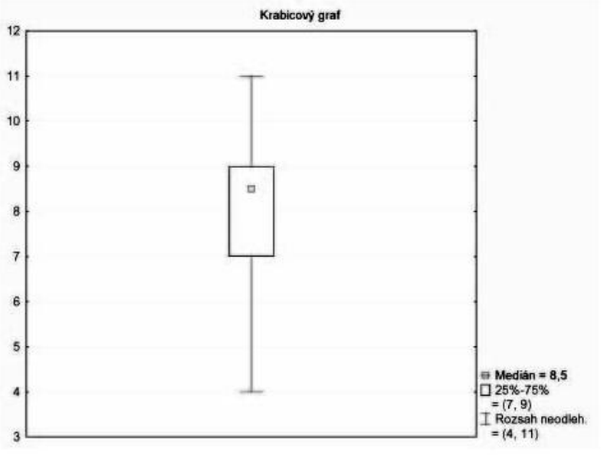
$$\bar{x} = \frac{3+4+4\cdot 7+4\cdot 8+6\cdot 9+3\cdot 10+11}{1+1+4+4+6+3+1} = \frac{162}{20} = 8{,}1.$$

Protože desátý člen v posloupnosti uspořádaných hodnot znaku je $x_{(10)} = 8$ a jedenáctý $x_{(11)} = 9$, je medián roven $\tilde{x} = \frac{8+9}{2} = 8{,}5$. Dolní kvartil je $x_{0,25} = \frac{x_{(5)}+x_{(6)}}{2} = 7$ a horní $x_{0,75} = \frac{x_{(15)}+x_{(16)}}{2} = 9$. Z definice rozptylu spočítáme

$$s_x^2 = \frac{5{,}1^2+4{,}1^2+4\cdot 1{,}1^2+4\cdot 0{,}1^2+6\cdot 0{,}9^2+3\cdot 1{,}9^2+2{,}9^2}{1+1+4+4+6+3+1} = \\ = 3{,}59.$$

Na následujících obrázcích je zobrazen příslušný histogram a krabicový diagram.





- *ordinálním*, kdy platí totéž jako předchozí, ale s přidaným uspořádáním (např. počet hvězdiček u hotelů v turistických průvodcích);
- *intervalovém*, kdy jde o číselné hodnoty, ale jde o porovnání velikostí, nikoliv absolutní hodnotu (např. u měření teplot je poloha nuly zpravidla dohodnuta, ale není podstatná);
- *poměrovém*, kdy máme pevně stanovené měřítko a nulu (např. většina fyzikálních nebo ekonomických veličin).

U nominálních typů znaků jsme schopni věcně interpretovat pouze rovnost $x_1 = x_2$, u ordinálních i nerovnost $x_1 < x_2$, případně $x_1 > x_2$, u intervalových navíc umíme posoudit rozdíl $x_1 - x_2$. U poměrových typů měřítek máme k dispozici rovnost, nerovnost, rozdíl i podíl x_1/x_2.

9.3. Třídění hodnot. V dalším budeme pracovat se *souborem hodnot* $x_1, x_2, \ldots, x_n$, které lze uspořádat (nejedná se tedy o hodnoty typy znaků nominálních) a které vznikly měřením na n statistických jednotkách, a uspořádáme je do *uspořádaného souboru hodnot*

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}. \tag{9.1}$$

Číslo n nazýváme *rozsah souboru*.

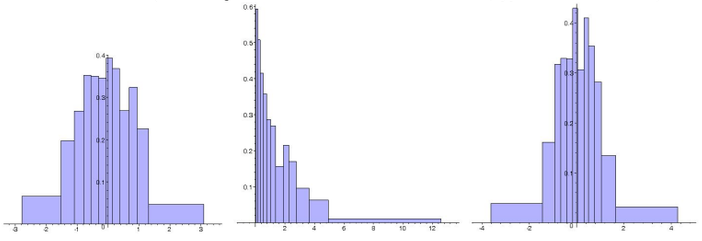
Pokud pracujeme s rozsáhlými soubory znaků, které ale připouští jen málo hodnot, je nejjednodušší uvádět pouze četnosti výskytu. Např. při průzkumu preferencí politických stran nebo u prezentace kvality hotelové sítě uvádíme u každé možné hodnoty počet jejích výskytů.

Pokud je možných hodnot mnoho (nebo dokonce připouštíme spojitě rozprostřené reálné hodnoty), dělíme často možný rozsah hodnot na vhodný počet intervalů a o statistickém znaku uvádíme četnost hodnot v daných intervalech. Intervalům se často říká *třídy* a počtu znaků ve třídě pak *třídní četnosti*. Používáme také *kumulativní četnosti* a *kumulativní třídní četnosti*, které pro danou třídu vznikají prostým součtem třídních četností s hodnotami nejvýše jako má ta daná.

Nejčastěji pak uvažujeme střed a_i dané třídy za hodnotu, která ji reprezentuje a hodnota $a_i n_i$, kde n_i je četnost výskytu této třídy představuje celkový příspěvek této třídy. Velmi často také místo četností zobrazujeme relativní četnosti a_i/n, resp. relativní kumulativní četnosti.

Graf, který na jedné ose vynáší intervaly jednotlivých tříd a nad nimi obdélníky s výškou rovnou četnosti se nazývá *histogram*. Obdobně se znázorňuje kumulativní četnost.

Na obrázku jsou histogramy souborů o rozsahu $n = 500$, které vznikly náhodným generováním dat s různými standardními rozděleními (časem jim budeme říkat normální, χ^2 a studentovo)



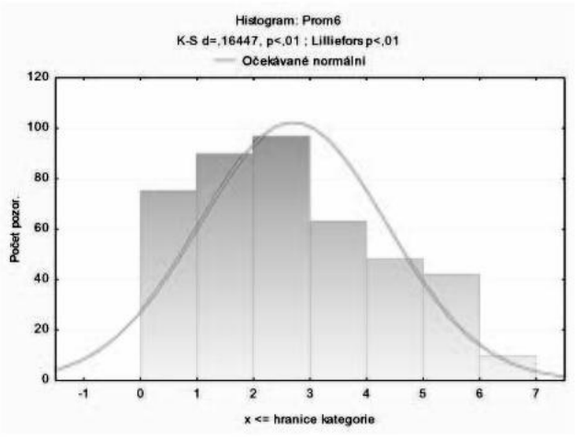
□

9.5. V daném rybníku se vylovilo 425 kaprů a u všech byly zjištěny jejich hmotnosti. Pak se vhodně zvolily hmotnostní intervaly a sestavila se následující tabulka četností:

Hmotnost (kg)	0-1	1-2	2-3	3-4	4-5	5-6	6-7
Střed třídy	0,5	1,5	2,5	3,5	4,5	5,5	6,5
Četnost	75	90	97	63	48	42	10

Načrtněte histogram, určete aritmetický, geometrický a harmonický průměr hmotnosti kaprů, dále určete medián, horní a dolní kvartil, modus, rozptyl, směrodatnou odchylku, variační koeficient a načrtněte příslušný krabicový diagram.

Řešení. Histogram má tvar



Z definic příslušných pojmů v části 9.4 přímo spočítáme aritmetický průměr $\bar{x} = 2{,}7\text{kg}$, geometrický průměr $\bar{x}^G = 2{,}1\text{kg}$, harmonický průměr $\bar{x}^H = 1{,}5\text{kg}$. Z definic v 9.5 je medián roven $\tilde{x} = x_{0,5} = 2{,}5\text{kg}$, dolní a horní kvartil $x_{0,25} = 1{,}5\text{kg}$ resp. $x_{0,75} = 3{,}5\text{kg}$ a pro modus platí $\hat{x} = 2{,}5\text{kg}$. Z definic v části 9.6 spočítáme rozptyl hmotnosti kaprů $s_x^2 = 2{,}7\text{kg}^2$, tj. směrodatná odchylka je $s_x = 1{,}7\text{kg}$, a variační koeficient $V_x = 0{,}6$. □

9.6. Dokažte, že entropie nabývá svého maxima, jsou-li hodnoty nominálního znaku rovnoměrně rozloženy, tj. četnost každé třídy je $n_i = 1$.

Řešení. Podle definice entropie 9.9 hledáme maximum funkce $H_X = -\sum_{i=1}^{n} p_i \ln p_i$ vzhledem k neznámým relativním četnostem $p_i = \frac{n_i}{n}$, které navíc splňují $\sum_{i=1}^{n} p_i = 1$. Jedná se tedy o klasickou úlohu hledání vázaného extrému, kterou můžeme vyřešit například pomocí Lagrangeových multiplikátorů. Příslušná Lagrangeova funkce je

$$L(p_1, \ldots, p_n, \lambda) = -\sum_{i=1}^{n} p_i \ln p_i + \lambda \left(\sum_{i=1}^{n} p_i - 1 \right).$$

9.4. Míry polohy statistických znaků. Chceme-li vyjádřit velikost hodnot, kolem kterých se jednotlivá pozorování znaků shromažďují používáme většinou pojmy z následující definice. Budeme teď pracovat se znaky poměrových (nebo případně intervalových) typů měřítek.

Uvažme (netříděný) soubor $(x_1, \ldots, x_n)$ hodnot měřeného znaku pro všechny zpracovávané statistické jednotky a nechť $n_1, \ldots, n_m$ jsou třídní četnosti m různých hodnot $a_1, \ldots, a_n$, nabývaných tímto souborem.

Průměry

Definice. *Aritmetický průměr* (často také jen *průměr*) je dán

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \sum_{j=1}^{m} n_j a_j.$$

Geometrický průměr je dán

$$\bar{x}^G = \sqrt[n]{x_1 x_2 \cdots x_n}$$

a má smysl pouze u kladných hodnot znaků. Harmonický průměr je dán

$$\bar{x}^H = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} \right)^{-1}$$

a je také definován jen pro kladné hodnoty znaků.

Aritmetický průměr je jediný z těchto průměrů, který je invariantní vůči afinním transformacím, tj. pro libovolné skaláry a, b platí

$$\overline{(a + b \cdot x)} = \frac{1}{n} \sum_{i=1}^{n} (a + b x_i) = a + b \sum_{i=1}^{n} x_i = a + b \cdot \bar{x}.$$

Aritmetický průměr je proto obzvlášť vhodný pro intervalové typy měřítek.

Logaritmus geometrického průměru je aritmetický průměr logaritmů znaků. Je obzvlášť vhodný pro znaky, které se kumulují multiplikativně, např. úrokové míry. Je-li totiž úroková míra v jednotlivých časových jednotkách $x_i\%$, bude za celé období výsledek takový, jakoby byla po celou dobu konstantní úroková míra $\bar{x}^G\%$.

Jako ilustraci tehdy rozvíjených metod jsme dokázali v odstavci 8.23 na straně 464, že je geometrický průměr vždy nejvýše tak velký jako aritmetický. Obdobně je tomu pro harmonický průměr a platí

$$\bar{x}^H \leq \bar{x}^G \leq \bar{x}.$$

9.5. Medián, kvartil, decil, percentil, ... Jiný způsob vyjádření míry, jakou hodnotu nabývají znaky, je najít pro číslo α mezi nulou a jedničkou takovou hodnotu x_α, aby $100\,\alpha\%$ hodnot znaku bylo nejvýše x_α a zbylé byly větší než x_α. Pokud takový znak není určen jednoznačně, volíme zpravidla průměr mezi dvěma extrémními možnými hodnotami.

Číslu x_α říkáme *α–kvantil*. Dosáhl-li tedy nějaký účastník soutěže výsledku, který jej řadí do $x_{1,00}$, neznamená to, že byl jistě lepší než všichni ostatní. Jen nebyl nikdo jiný ještě lepší než on.

Nejobvyklejší hodnoty x_α jsou:

- *medián* (často také výběrový medián) definovaný vztahem

$$\tilde{x} = x_{0,50} = \begin{cases} x_{((n+1)/2)} & \text{pro liché } n \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) & \text{pro sudé } n \end{cases},$$

Pro její parciální derivace platí $\frac{\partial L}{\partial p_i} = -\ln p_i - 1 + \lambda$, a proto je její stacionární bod určen rovnicemi $p_i = e^{\lambda-1}$ pro všechna $i = 1, \ldots, n$. Navíc víme, že součet relativních četností p_i je roven jedné. To znamená $ne^{\lambda-1} = 1$ a odtud $\lambda = 1 - \ln n$. Dosazením zřejmě $p_i = \frac{1}{n}$. □

9.7. Následující grafy udávají četnosti možných bodových zisků studentů předmětu MB104 na Fakultě informaticky Masarykovy univerzity v roce 2012. Kumulativní graf je uváděn s „prohozenými“ osami oproti předchozímu příkladu.

Četnosti jednotlivých bodových zisků jsou uvedeny v následující tabulce:

Body	Počet studentů
20.5	1
20	1
19	2
18.5	1
18	2
17.5	3
17	2
16.5	4
16	3
15.5	5
15	7
14.5	6
14	14
13.5	21
13	21
12.5	19
12	17
11.5	18
11	31
10.5	22
10	53

Body	Počet studentů
9.5	9
9	9
8.5	13
8	8
7.5	13
7	4
6.5	7
6	4
5.5	8
5	7
4.5	9
4	5
3.5	7
3	8
2.5	8
2	14
1.5	8
1	2
0.5	6
0	9

Tomu potom odpovídá následující histogram:

kde $x_{(k)}$ představuje hodnotu v uspořádaném souboru hodnot (9.1)

- *dolní a horní kvartil* $Q_1 = x_{0,25}$ a $Q_3 = x_{0,75}$;
- *p-tý kvantil* (též *výběrový kvantil* nebo *percentil*) x_p, kde $0 < p < 1$ (zpravidla zadaný na dvě desetinná místa).

Lze se setkat také s hodnotou *modus*, která udává hodnotu $\hat{x}$ znaku s největší četností v souboru x.

Aritmetický průměr, medián a modus představují jakési očekávatelné hodnoty znaků. Průměr u znaku podílového typu, medián u poměrového a modus u ordinálního nebo nominálního.

Všimněme si, že všechny α–kvantily hodnot v intervalových měřítcích jsou invariantní vzhledem k afinním transformacím hodnot (promyslete si podrobně!).

9.6. Míry variability statistických znaků. Rozumným požadavkem na jakoukoliv míru variability souboru hodnot znaků $x \in \mathbb{R}^n$ je její invariance vůči konstantním posunutím. V euklidovském prostoru $\mathbb{R}^n$ má tuto vlastnost standardní vzdálenost bodů a nezávislý na posunutí o konstantní hodnotu je i výběrový průměr. Proto volíme následující

ROZPTYL A SMĚRODATNÁ ODCHYLKA

Definice. *Rozptyl* souboru znaků x je definován vztahem

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_i)^2.$$

Směrodatná odchylka s_x je dána jako odmocnina z výběrového rozptylu.

Často se v literatuře také pro rozptyl používá název *střední kvadratická odchylka*.

Variabilita statistických znaků by neměla záviset na konstantním posunutí všech hodnot. Při naší definici jsme proto vyšli z toho, že jak standardní vzdálenost bodů v $\mathbb{R}^n$ tak výběrový průměr jsou vůči posunutím o konstantní hodnotu invariantní, bude proto skutečně i pro neuspořádaný soubor znaků

$$y = (x_1 + c, x_2 + c, \ldots, x_n + c)$$

vždy platit také $s_y = s_x$.

Někdy se místo naší hodnoty s_x používá tzv. *výběrový rozptyl*, který se odlišuje jen tím, že se ve jmenovateli zlomku používá $(n - 1)$, důvod uvidíme později.

V případě třídních četností n_j hodnot a_j pro m tříd dává stejný výraz hodnotu rozptylu

$$s_x^2 = \frac{1}{n}\sum_{j=1}^{m} n_j(a_j - \bar{x})^2,$$

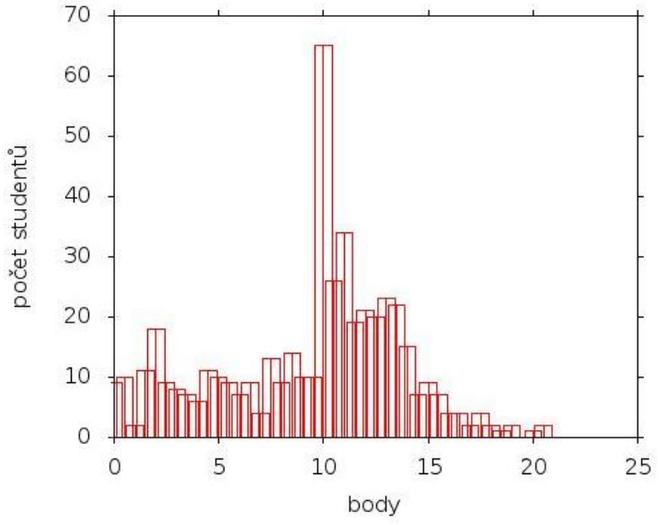
ale v praxi se doporučuje používat tzv. Shepardovu korekci, která S_x^2 zmenší o $h^2/12$, kde h je šířka stejných intervalů definujících třídy hodnot.

Dále se ještě můžeme potkat s tzv. *rozpětím výběru*

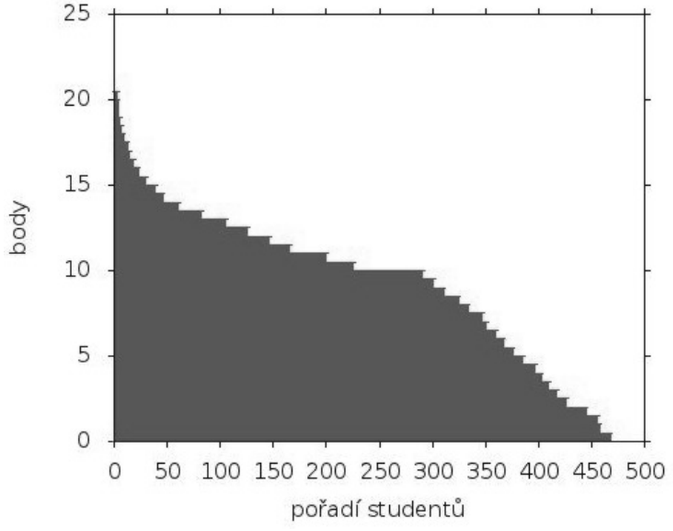
$$R = x_{(n)} - x_{(1)}$$

a *kvartilovým rozpětím výběru*

$$Q = Q_3 - Q_1.$$



Histogram jsme obdrželi z informačního systému Masarykovy univerzity. Vidíme, že je zvolen poněkud netradiční způsob zobrazování, kdy danému bodovému zisku odpovídá „dvojitý obdélníček“. Je na vkusu každého čtenáře, jaký způsob výpisu dat zvolí (je možno některé hodnoty počítat do jedné, čímž snížíme počet obdélníčků, nebo používat tenčí obdélníčky).



Snadno si všimneme, že modusem bodových hodnot je číslo 10, což byla shodou okolností bodová hranice zaručující absolvování předmětu. Průměr získaných bodů je 9,48.

9.8. Uveďme ještě sloupcové diagramy bodových zisků studentů předmětu MB101 v podzimním semestru 2010 (první semestr studia) a to jednak všech účastníků předmětu a poté studentů, kteří úspěšně ukončili bakalářské studium.

Používá se také tzv. *průměrná odchylka*, která je dána průměrnou vzdáleností hodnot od mediánu

$$D_x = \frac{1}{n}\sum_{i=1}^{n}|x_i - \tilde{x}|.$$

Následující věta podává zdůvodnění, proč tyto míry variability volíme:

Věta. *Funkce* $S(t) = (1/n)\sum_{i=1}^{n}(x_i - t)^2$ *nabývá svého minima pro* $t = \bar{x}$*, tj. pro výběrový průměr.*

Funkce $D(t) = (1/n)\sum_{i=1}^{n}|x_i - t|$ *nabývá svého minima pro* $t = \tilde{x}$*, tj. pro medián.*

Důkaz. Protože je součet vzdáleností všech hodnot od výběrového průměru nulový, dostáváme přímým výpočtem

$$\begin{aligned}\sum_{i=1}^{n}(x_i - t)^2 &= \sum_{i=1}^{n}\big((x_i - \bar{x})^2 + (\bar{x} - t)^2 - 2(x_i - \bar{x})(\bar{x} - t)\big)\\ &= n(\bar{x} - t)^2 + \sum_{i=1}^{n}(x_i - \bar{x})^2,\end{aligned}$$

což ověřuje první tvrzení.

U druhého si musíme dát pozor na definici mediánu. Součet si za tím účelem přeskládáme tak, abychom vždy postupně sčítali první s posledním sčítancem, pak druhý s předposledním atd. V prvním případě tedy jde o výraz $|x_{(1)} - t| + |x_{(n)} - t|$, a ten bude roven vzdálenosti $x_{(n)} - x_{(1)}$, pokud bude t uvnitř rozsahu hodnost, a bude ještě vetší jinak. Další dvojice v součtu nám stejně dá $x_{(n-1)} - x_{(2)}$, pokud bude $x_{(2)} \le t \le x_{(n-1)}$ a bude větší jinak. Postupně tedy požadavek na minimalizaci součtu povede právě na $t = \tilde{x}$. □

V praxi potřebujeme poměřovat variabilitu různých souborů hodnot znaků různých statistických jednotek. Pro tento účel je vhodné relativizovat měřítko a používáme proto tzv. *variační koeficient* daného souboru x

$$V_X = \frac{\sqrt{s_x^2}}{|\bar{x}|}.$$

Tuto relativní míru variability lze také chápat v procentech směrodatné odchylky ve vztahu k výběrovému průměru $\bar{x}$.

9.7. Šikmost rozložení hodnot znaků. Pokud jsou rozloženy znaky našeho souboru naprosto symetricky kolem výběrového průměru, bude zejména platit

$$\bar{x} = \tilde{x}$$

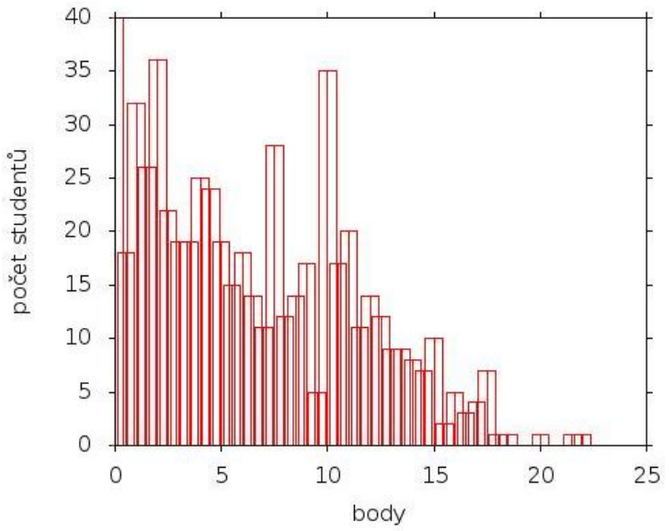
Často ale potkáváme rozložení hodnot splňujících

$$\bar{x} > \tilde{x},$$

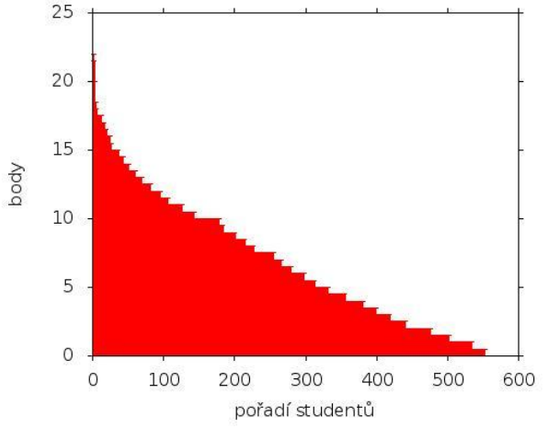
např. to je běžné u rozložení mezd v populaci. Docela užitečnou charakteristikou v tomto směru je tzv. Pearsonův koeficient, který je dán vztahem

$$\beta = 3\frac{\bar{x} - \tilde{x}}{S_x}$$

a můžeme si z něho udělat představu o relativní míře (absolutní hodnota β) i charakteru zešikmení (znaménko). Zejména si všimněme, že směrodatná odchylka je vždy kladná, takže již znaménko nám ukazuje, kterým směrem k zešikmení dochází.



Výsledky opět můžeme zachytit i alternativně:



A nyní grafy bodových zisků účastníků, kteří dále úspěšně pokračovali ve studiu.

Kvantilové koeficienty šikmosti

Podrobnější informaci v tomto směru dávají tzv. *kvantilové koeficienty šikmosti*

$$\beta_p = \frac{x_{1-p} + x_p}{x_{1-p} - x_p},$$

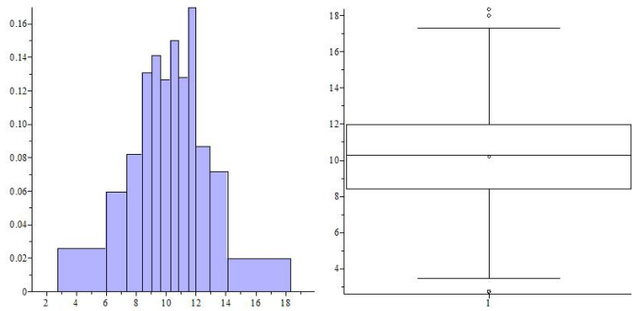
pro každé $0 < p < 1$. Jejich význam je zřejmý, když čitatel zlomku vyjádříme jako $(x_{1-p} - \tilde{x}) - (\tilde{x} - x_p)$.

Speciálně dostáváme tzv. *kvartilový koeficient šikmosti* při volbě $p = 0{,}25$.

9.8. Diagramy. Pro rychlé vstřebávání složitěji strukturovaných informací je člověk skvěle vybaven zrakově. Proto se pro zobrazení statistiky jednotlivých znaků nebo jejich korelací používá mnoho standardizovaných nástrojů. Jedním z nich jsou tzv. *krabicové diagramy*.

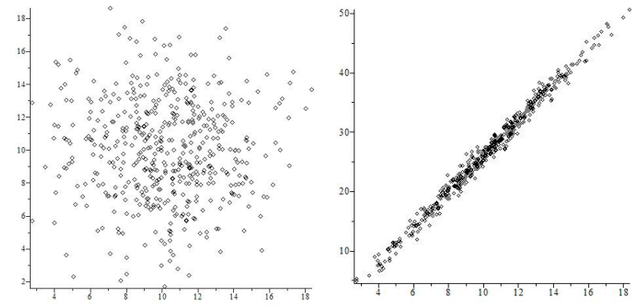
Krabicový diagram

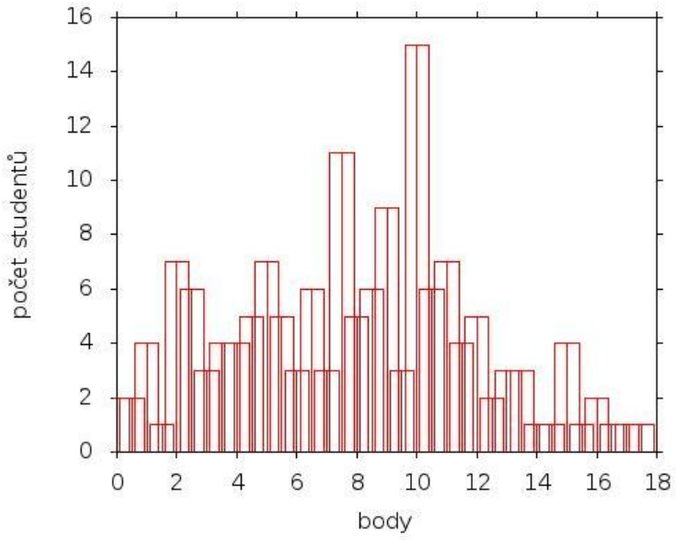
Na obrázku je zobrazen histogram a krabicový diagram stejného souboru hodnot (normální rozdělení s průměrem 10 a rozptylem 3, $n = 500$).

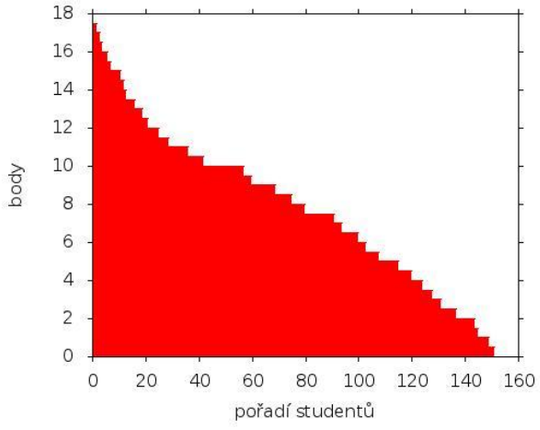


Střední linka je medián, kraje boxu jsou kvartily, „packy“ ukazují 1,5 kvartilového rozsahu, ne však víc než kraje rozsahu výběru, případné hodnoty mimo jsou přímo naznačeny body.

Běžné zobrazovací nástroje nám umožňují dobře vidět případné závislosti dvou výběrů zjištěných znaků. Např. na levém obrázku níže jsou za souřadnice voleny hodnoty ze dvou nezávislých normálních rozdělení s průměrem 10 a rozptylem 3. Na pravém obrázku je první souřadnice ze stejných dat, druhá je z první dána vztahem $y = 3x + 4$, ale je navíc zatížená malou náhodnou chybou.







Vidíme, že modus zisku bodů v prvním případě je 0, ve druhém případě je to opět deset. Rozložení bodových zisků se blíží rozložení bodových zisků z předmětu MB104, který je zařazen ve čtvrtém semestru studia.

9.9. Auto jelo z Brna do Prahy rychlostí 160 km/h, z Prahy do Brna rychlostí 120 km/h. Jaké průměrné rychlosti na trase dosáhlo?

Řešení. Toto je základní příklad, kde je použití aritmetického průměru nevhodné. Na průměrnou rychlost totiž klademe požadavek, aby auto jedoucí touto rychlostí strávilo na trase stejnou dobu. Označíme-li d vzdálenost obou měst v kilometrech, v_p průměrnou rychlost tak

$$\frac{d}{160}+\frac{d}{120}=\frac{2d}{v_p},$$

odkud

$$v_p=\frac{2}{\frac{1}{160}+\frac{1}{120}}\doteq 137{,}14.$$

9.9. Entropie. Variabilitu potřebujeme vyjadřovat i u nominálních typů znaků, např. ve statistické fyzice nebo teorii informace. K dispozici máme jen třídní četnosti a můžeme tedy použít princip klasické pravděpodobnosti (viz čtvrtá část první kapitoly), kdy relativní četnost i-té třídy, $p_i=\frac{n_i}{n}$, vnímáme jako pravděpodobnost, že náhodně vybraný prvek bude v této třídě.

Rozptyl poměrových hodnot znaku, u kterého máme vyjádřeny třídní četnosti n_j, byl v odstavci 9.6 vyjádřen vztahem

$$s_x^2=\sum_{j=1}^{m}\frac{n_j}{n}(a_j-\bar{x})^2=\sum_{j=1}^{m}p_j(a_j-\bar{x})^2,$$

kde p_j označuje (klasickou) pravděpodobnost, že hodnota znaku bude v j-té třídě. Jde tedy o vážený průměr přepočtených hodnot znaků, kde je hodnota $F(a_j)=(a_j-\bar{x})^2$ vstupuje s vahou p_j.

Variabilitu hodnot znaků nominálního typu budeme vyjadřovat podobným výrazem, označíme ho H_X. Nemáme sice k dispozici žádné číselné hodnoty a_j pro pořadové indexy j, můžeme se ale zajímat o funkce F závisející na relativních četnostech p_j, tj. zkusíme pro datový soubor x definovat

$$H_X=\sum_{i=1}^{n}p_iF(p_i),$$

kde F je zatím neznámá funkce.

Pokud znak nabývá právě jedné hodnoty, tj. pokud $p_k=1$ pro nějaké k a všechna ostatní $p_j=0$, pak budeme jistě říkat, že variabilita je nulová. Je tedy v každém případě $F(1)=0$.

Dále budeme požadovat, aby H_X měla následující vlastnost. Pokud je zkoumaný soubor znaků Z tvořen dvojicemi znaků ze souborů X a Y (např. můžeme na statistických jednotkách-osobách sledovat barvu očí a barvu vlasů), je rozumné, aby variabilita znaků z byla součtem variabilit jednotlivých znaků, tj. požadujeme v takovém případě $H_Z=H_X+H_Y$.

Známe relativní třídní četnosti p_i pro znaky v souboru X a q_j pro znaky souboru Y. Relativní třídní četnosti pro Z jsou

$$r_{ij}=\frac{n_im_j}{nm}=p_iq_j$$

a požadujeme tedy rovnost (rozsahy součtů jsou zřejmé z kontextu)

$$\sum_{i,j}p_iq_jF(p_iq_j)=\sum_i p_iF(p_i)+\sum_j q_jF(q_j).$$

Díky tomu, že p_i a q_j jsou relativní četnosti a tedy dávají v součtu 1, můžeme pravou stranu rovnosti přepsat jako

$$\Big(\sum_j q_j\Big)\Big(\sum_i p_iF(p_i)\Big)+\Big(\sum_i p_i\Big)\Big(\sum_j q_jF(q_j)\Big)$$

a dostáváme vztah

$$\sum_{i,j}p_iq_jF(p_iq_j)=\sum_{i,j}p_iq_j\big(F(p_i)+F(q_j)\big).$$

Je zřejmé, že tomuto požadavku vyhovuje jakýkoliv konstantní násobek logaritmu při kterémkoliv pevně zvoleném základu $a>1$ (a lze ukázat, že jiná spojitá řešení F neexistují).

Poněvadž je $p_i\le 1$, je jistě $\ln p_i\le 0$. My však chceme variabilitu nezápornou, zvolíme proto za funkci F logaritmickou funkci s násobkem -1. Taková volba také automaticky splňuje náš požadavek $F(1)=0$.

Průměrná rychlost je tedy dána harmonickým průměrem (viz 9.3 průměrovaných rychlostí

□

B. Vizualizace vícerozměrných dat.

V předchozích příkladech jsme se věnovali zobrazování jednoho znaku měřeného u více objektů (získané body studentů). Grafická vizualizace dat pomáhá k lepší představě o datech. Jak ale postupovat, pokud u některých, řekněme n objektů, měříme nějakých p znaků, $p \geq 3$. Tato měření není možné znázornit způsoby, které jsme se již naučili. Jednou z možných metod, je tzv. *metoda hlavních komponent*. V této metodě využijeme pojmu vlastního vektoru a vlastních čísel (viz 2.46) výběrové varianční matice (viz 9.38). Zaveďme následující označení:

- náhodné vektory měření $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})^T$, $i = 1, \ldots, n$,
- průměr j-tého znaku $m_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, $j = 1, \ldots, p$,
- rozptyl j-tého znaku $s_j = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - m_j)^2$, $j = 1, \ldots, p$,
- vektor průměrů $\mathbf{m} = (m_1, \ldots, m_p)$,
- výběrová varianční matice $\frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T$ (všimněme si, že každý sčítanec v předchozí sumě je maticí rozměrů $p \times p$).

Varianční matice je symetrická, tudíž má všechna vlastní čísla reálná a její vlastní vektory jsou navzájem kolmé. Volíme-li navíc vlastní vektory jednotkové, pak z toho vyplývá, že vlastní hodnota příslušná nějakému vlastnímu vektoru varianční matice dává rozptyl (velikosti) průmětu daných dat do tohoto směru (promítáme v p-rozměrném prostoru). Cílem této metody je nalézt směr (v p-rozměrném prostoru znaků), pro který je rozptyl průmětů daných dat do něj největší. Tento směr tedy odpovídá tomu vlastnímu vektoru varianční matice, který odpovídá největší vlastní hodnotě. Lineární kombinace daná složkami tohoto vektoru se nazývá 1. hlavní komponenta. Velikost průmětu daných dat do tohoto směru relativně dobře odhaduje data (hlavní komponentu lze chápat jako jeden znak, který nahrazuje p znaků, jde tedy o náhodný vektor o n položkách). Pokud od dat odečteme tento průmět a opět uvážíme směr největší variability takto pozměněných dat, dostáváme 2. hlavní komponentu a opakováním tohoto postupu dostáváme další hlavní komponenty. Směr největší variability je ovšem vlastní vektor varianční matice odpovídající největšímu vlastnímu číslu (čtenář si laskavě rozmyslí). Směry dalších hlavních

Entropie

Míru variability znaků v nominálním měřítku vyjadřujeme pomocí *entropie*. Je dána vztahem

$$H_X = -\sum_{i=1}^{k} \frac{n_i}{n} \ln\left(\frac{n_i}{n}\right),$$

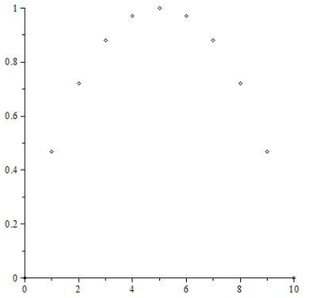
kde k je počet tříd ve výběru. Kromě přirozeného logaritmu se často také setkáváme (např. teorii informace) se stejným vztahem ale s logaritmem při základu 2.

Často se také místo H_X pracuje s veličinou

$$e^{H_X} = \prod_i p_i^{-p_i},$$

případně totéž s jiným zvoleným základem pro logaritmus.

V tomto tvaru se pěkně spočítá, že pro výběr X s k stejně velkými třídními četnostmi je $e^{H_X} = \left((\frac{1}{k})^{-\frac{1}{k}}\right)^k = k$, nezávisle na velikosti výběru. Na obrázku jsou vyneseny entropie y při základu 2 pro výskyt písmen a a b v desetipísmenných slovech s písmeny a a b, kde x je počet výskytů písmene b.



Všimněme si, že pro shodný výskyt, tj. pro pět písmen b, vyjde maximální entropie 1 a skutečně je $2^1 = 2$.

2. Pravděpodobnost

Před dalším čtením lze čtenářům vřele doporučit zopakování obsahu čtvrté části první kapitoly (tj. odstavce začínající na straně 17). Tehdy jsme pracovali převážně s tzv. klasickou konečnou pravděpodobností a zavedli jsme základy formalismu, který nyní rozšíříme. Hlavní změnou bude, že náš základní prostor Ω už nebude obecně obsahovat jen konečně mnoho prvků (ve skutečnosti nemusí být ani spočetný). Připomeňme, že v našich úvahách o tzv. geometrické pravděpodobnosti na konci čtvrté části první kapitoly jsme potřebovali jako základní prostor pro popis jevu vhodnou část euklidovského prostoru a jevy pak byly vhodně vybrané podmnožiny. Tedy samozřejmě samé nespočetné množiny.

Začneme jednoduchým, stále ještě diskrétním, ale nekonečným příkladem, ke kterému se ve výkladu budeme občas vracet.

9.10. Proč nekonečné množiny jevů? Představme si experiment, ve kterém opakovaně házíme mincí dokud nepadne líc. Ptáme se, jaká je pravděpodobnost, že budeme házet alespoň 3–krát nebo právě 35–krát nebo nejvýš 10–krát apod.

komponent pak odpovídají dalším vlastním vektorům varianční matice (seřazenými podle velikosti vlastních hodnot, které jim přísluší).

9.10. Určete 1. hlavní komponentu následujících jednoduchých dat a vektor, který jejím použitím nahrazuje naměřená data. U pěti osob byla změřena výška, délka malíčku a délka ukazováčku s výsledky zaznamenanými v tabulce (výsledky jsou v centimetrech).

Řešení.

	Martin	Michal	Matěj	Honza	Markéta
ukazováček	9	11	8	8	8
malíček	7,5	8	6,3,	6	6,5
výška	186	187	173	174	167

.

Vektory pozorovaných hodnot jsou: $\mathbf{x}_1 = (9; 7{,}5; 186)$, $\mathbf{x}_2 = (11; 8; 187)$, $\mathbf{x}_3 = (8; 6; 173)$, $\mathbf{x}_4 = (8; 6; 174)$, $\mathbf{x}_5 = (8; 6{,}5, 167)$. Varianční matice těchto vektorů jsou postupně

$$\begin{pmatrix} 0{,}04 & 0{,}14 & 1{,}72 \\ 0{,}14 & 0{,}49 & 6{,}02 \\ 1{,}72 & 6{,}02 & 73{,}96 \end{pmatrix}, \quad \begin{pmatrix} 4{,}840 & 2{,}64 & 21{,}12 \\ 2{,}64 & 1{,}44 & 11{,}52 \\ 21{,}12 & 11{,}52 & 92{,}16 \end{pmatrix},$$

$$\begin{pmatrix} 0{,}641 & 0{,}640 & 3{,}521 \\ 0{,}640 & 0{,}640 & 3{,}52 \\ 3{,}521 & 3{,}52 & 19{,}36 \end{pmatrix}, \quad \begin{pmatrix} 0{,}641 & 0{,}640 & 2{,}721 \\ 0{,}640 & 0{,}640 & 2{,}72 \\ 2{,}721 & 2{,}72 & 11{,}56 \end{pmatrix},$$

$$\begin{pmatrix} 0{,}641 & 0{,}240 & 8{,}321 \\ 0{,}240 & 0{,}09 & 3{,}12 \\ 8{,}32 & 3{,}12 & 108{,}16 \end{pmatrix}.$$

Výběrovou varianční matice je pak čtvrtina ze součtu těchto matic, tedy

$$S = \begin{pmatrix} 1{,}70 & 1{,}075 & 9{,}35 \\ 1{,}075 & 0{,}825 & 6{,}725 \\ 9{,}35 & 6{,}725 & 76{,}30 \end{pmatrix}$$

Vlastní hodnoty matice S jsou přibližně 2,7, 312,2 a 0,38. Jednotkový vlastní vektor odpovídající největší z nich pak cirka (0,122; 0,09; 0,989). První hlavní komponenta je tedy (185 5; 186,8; 172,4; 173,4; 166,5), tedy se příliš neliší od výšky zkoumaných osob. □

9.11. Žáci jedné třídy dosáhli následujících známek v různých předmětech:

Elementární jevy bychom tedy mohli uvažovat ve tvaru $\omega_k \in \mathbb{N}_{\geq 1} \cup \{\infty\}$, které slovně vyjadřujeme „líc padne poprvé právě v k–tém hodu“. Všimněme si, že jsme přidali $k = \infty$, protože formálně nemůžeme vyloučit, že budou vždycky padat pouze ruby mince.

Zjevně můžeme takový problém dobře zvládat, když vyjdeme z klasické pravděpodobnosti 0,5 pro obě možné strany mince při jednom hodu, nemůžeme ale v abstraktním modelu omezit celkový počet hodů nějakým pevným přirozeným číslem N. Na druhé straně, očekávaná pravděpodobnost, že padne ve všech prvních $(k-1)$ pokusech vždy rub v $n \geq k$ pokusech celkem, je dána zlomkem

$$\frac{2^{n-k}}{2^n} = 2^{-k},$$

kde v čitateli je počet možností příznivých z n nezávislých hodů (tj. možností jak rozestavit libovolně dvě hodnoty do $n-k$ zbývajících pozic) a ve jmenovateli je počet všech možností výsledků. Podle očekávání tato pravděpodobnost nezávisí na zvoleném n a platí $\sum_{k=1}^{\infty} 2^{-k} = 1$. Musí být proto pravděpodobnost neustálého opakování rubu nulová.

Můžeme tedy nyní zavést skutečně pravděpodobnost na základní prostoru Ω s elementárními jevy ω_k, kterým přiřazujeme pravděpodobnost 2^{-k}. Dostaneme tak pravděpodobnostní prostor ve smyslu následujících definic.

K tomuto jednoduchému ilustračnímu příkladu se ještě budeme vracet.

9.11. Jevová pole. Budeme pracovat s neprázdnou pevně zvolenou množinou Ω ve které se budou odehrávat všechny výsledky a kterou nazýváme *základní prostor*. Prvky $\omega \in \Omega$ představují jednotlivé *možné výsledky*. V pravděpodobnostních modelech ale nemusíme připouštět všechny možné podmnožiny coby uvažované jevy. Zejména jednotlivé prvky ω nemusí být mezi jevy. Požadujeme ale, aby uvažované podmnožiny splňovaly axiomy tzv. *σ–algeber*.

Níže uvedené axiomy jsou vybrány z větší sady přirozených požadavků v minimální podobě. První vychází z představy, že určitě budeme chtít připustit jev jistý. Druhý je vynucen požadavkem, že chceme vždy mít možnost negovat výskyt jevu, třetí potřebou zkoumat výskyt alespoň jednoho z dané spočetné množiny jevů (např. v případech podobných tomu v předcházejícím odstavci, kdy sice víme, že nikdo nehodí mincí nekonečně krát, nicméně nemůžeme předem omezit počet hodů).

σ–ALGEBRY PODMNOŽIN

Systém podmnožin $\mathcal{A}$ základního prostoru se nazývá *jevové pole* a jeho prvky se nazývají *jevy*, jestliže platí

- $\Omega \in \mathcal{A}$, tj. základní prostor, je jevem,
- je-li $A, B \in \mathcal{A}$, pak $A \setminus B \in \mathcal{A}$, tj. pro každé dva jevy je jevem i jejich množinový rozdíl,
- je-li $A_i \in \mathcal{A}$, $i \in I$, nejvýše spočetný systém jevů, pak také jejich sjednocení je jevem, tj. $\cup_{i \in I} A_i \in \mathcal{A}$.

Jako obvykle, ze základních axiomů hned vyplývají jednoduché důsledky, které popisují další (intuitivně požadované) vlastnosti ve formě matematických vět. Čtenář by si měl promyslet, že obě vlastnosti skutečně platí.

Žák číslo	Matika	Fyzika	Dějepis	ČJ	Tělocvik
1	1	1	2	2	1
2	1	3	1	1	1
3	2	1	1	1	1
4	2	2	2	2	1
5	1	1	3	2	1
6	2	1	2	1	2
7	3	3	2	2	1
8	3	2	1	1	1
9	4	3	2	3	1
10	2	3	1	2	1

Určete první hlavní komponentu těchto dat a vektor dat, který jejím použitím nahrazuje původní data.

Řešení. Vektory pozorování jsou $\mathbf{x}_1 = (1,1,2,2,1), \ldots, \mathbf{x}_{10} = (2,3,1,2,1)$, jim odpovídající varianční matice pak

$$\begin{pmatrix} 1{,}21 & 1{,}10 & -0{,}330 & -0{,}330 & 0{,}110 \\ 1{,}10 & 1, & -0{,}300 & -0{,}300 & 0{,}100 \\ -0{,}330 & -0{,}300 & 0{,}0900 & 0{,}0900 & -0{,}0300 \\ -0{,}330 & -0{,}300 & 0{,}0900 & 0{,}0900 & -0{,}0300 \\ 0{,}110 & 0{,}100 & -0{,}0300 & -0{,}0300 & 0{,}0100 \end{pmatrix}, \ldots,$$

$$\begin{pmatrix} 0{,}0100 & -0{,}100 & 0{,}0701 & -0{,}0300 & 0{,}0100 \\ -0{,}100 & 1, & -0{,}700 & 0{,}300 & -0{,}100 \\ 0{,}0701 & -0{,}700 & 0{,}490 & -0{,}210 & 0{,}0701 \\ -0{,}0300 & 0{,}300 & -0{,}210 & 0{,}0900 & -0{,}0300 \\ 0{,}0100 & -0{,}100 & 0{,}0701 & -0{,}0300 & 0{,}0100 \end{pmatrix}$$

Výběrová varianční matice je pak

$$\begin{pmatrix} 0{,}99 & 0{,}44 & -0{,}078 & 0{,}26 & -0{,}01 \\ 0{,}44 & 0{,}89 & -0{,}22 & 0{,}22 & -0{,}11 \\ -0{,}078 & -0{,}22 & 0{,}45 & 0{,}23 & 0{,}03 \\ 0{,}26 & 0{,}22 & 0{,}23 & 0{,}45 & -0{,}078 \\ -0{,}01 & -0{,}11 & 0{,}033 & -0{,}0778 & 0{,}100 \end{pmatrix},$$

její dominantní vlastní hodnota je pak cca 13,68 a jí příslušný jednotkový vlastní vektor je přibližně (0,70; 0,65; −0,13; 0,28; −0,07). Hlavní komponenta je tedy (1,58; 2,73; 2,13; 2,93; 1,45; 1,93; 4,28; 3,48; 5,26; 3,71) □

Další možnou metodou vizualizace vícerozměrných dat je tzv. shluková analýza, tou se ale zabývat nebudeme.

C. Klasická a podmíněná pravděpodobnost

V první kapitole jsme se již seznámili s klasickou pravděpodobností, viz 1.13. Pro připomenutí si uveďme některé komplikovanější příklady.

9.12. Alešovi zbylo 2500 Kč z pořádání tábora. Aleš není žádný ňouma: 50 Kč přidal z kasičky a rozhodl se jít hrát ruletu na automaty. Aleš sází pouze na barvu. Pravděpodobnost výhry při sázce na barvu

Komplementy a průniky

Budeme využívat následující důsledky a terminologii:

- Komplement $A^c = \Omega \setminus A$ jevu A je jevem, který nazýváme *opačný jev* k jevu A.
- Průnik dvou jevů opět jevem, protože pro každé dvě podmnožiny $A, B \subset \Omega$ platí

$$A \setminus (\Omega \setminus B) = A \cap B.$$

Hovoříme přitom o *současném nastoupení jevů* A a B

Jevové pole je tedy systém podmnožin základního prostoru uzavřený na konečné průniky, spočetná sjednocení a množinové rozdíly.

Jednotlivé množiny $A \in \mathcal{A}$ nazýváme *náhodné jevy* (vzhledem k $\mathcal{A}$).

9.12. Pravděpodobnostní prostor. Teď popíšeme, co bude v našem matematickém modelu pravděpodobnost. Nejdříve ale ještě připomeneme názvosloví užívané už v první kapitole.

Terminologie

Používáme následující názvy týkající se jevů:

- celý základní prostor Ω se nazývá *jistý jev*, prázdná podmnožina $\emptyset \in \mathcal{A}$ se nazývá *nemožný jev*;
- jednoprvkové podmnožiny $\{\omega\} \in \Omega$ se nazývají *elementární jevy*;
- průnik jevů $\cap_{i\in I} A_i$ odpovídá *společnému nastoupení jevů* A_i, $i \in I$;
- sjednocení jevů $\cup_{i\in I} A_i$ odpovídá *nastoupení alespoň jednoho z jevů* A_i, $i \in I$;
- je-li $A \cap B = \emptyset$, pak se jevy $A, B \in \mathcal{A}$ nazývají *neslučitelné*,
- je-li $A \subset B$, pak říkáme, že jev A má za *důsledek* jev B;
- je-li $A \in \mathcal{A}$, pak se jev $B = \Omega \setminus A$ nazývá *opačný jev k jevu* A, píšeme $B = A^c$.

Hned v prvním odstavci této části jsme viděli příklad pravděpodobnosti definované na nekonečné množině elementárních jevů. Obecně budeme pravděpodobnost chápat takto:

Pravděpodobnost

Definice. *Pravděpodobnostní prostor* je jevové pole $\mathcal{A}$ podmnožin základního prostoru Ω, na kterém je definována skalární funkce $P : \mathcal{A} \to \mathbb{R}$ s následujícími vlastnosti:

- P je nezáporná, tj. $P(A) \geq 0$ pro všechny jevy A,
- P je spočetně aditivní, tj. $P(\cup_{i\in I} A_i) = \sum_{i\in I} P(A_i)$, pro každý nejvýše spočetný systém po dvou neslučitelných jevů,
- pravděpodobnost jistého jevu je 1.

Funkci P říkáme *pravděpodobnost* na jevovém poli $(\Omega, \mathcal{A})$.

Z definice okamžitě vidíme, že pro opačné jevy platí

$$P(A^c) = 1 - P(A).$$

Podobně zůstávají v platnosti důkazy, které jsme o sčítání pravděpodobností odvodili pro konečné systémy (protože vztahy stejně vždy obsahovaly pouze konečně mnoho množin – promyslete si podrobněji!) Zejména tedy platí pro libovolnou množinu k

je 18/37. Začíná sázet na 10 Kč a pokud prohraje, v další sázce vsadí dvojnásobek toho, co v předchozí (pokud na to ještě má, pokud ne, tak končí s hrou – byť by měl ještě peníze na nějakou menší sázku). Pokud nějakou sázku vyhraje, v následující sázce hraje opět o 10 Kč. Jaká je pravděpodobnost, že při tomto postupu vyhraje dalších 2550 Kč? (jakmile bude 2550 Kč v plusu, tak končí)

Řešení. Nejprve spočítejme, kolikrát po sobě může Aleš prohrát. Začíná-li s 10 Kč, tak na n vsazení potřebuje

$$10{+}20{+}\cdots{+}10{\cdot}2^{n-1} = 10{\cdot}\left(\sum_{i=0}^{n-1} 2^i\right) = 10{\cdot}\left(\frac{2^n-1}{2-1}\right) = 10{\cdot}(2^n{-}1).$$

Jak snadno nahlédneme, číslo 2550 je tvaru $10(2^n-1)$ a to pro $n=8$. Aleš tedy může sázet osmkrát po sobě bez ohledu na výsledek sázky, na devět sázek by potřeboval již $10(2^9-1) = 5110$ Kč a to v průběhu hry nikdy mít nebude (jakmile bude mít 5100 Kč, tak končí). Aby tedy jeho hra skončila neúspěchem, musel by prohrát osmkrát v řadě. Pravděpodobnost prohry při jedné sázce je 19/37, pravděpodobnost prohry v osmi po sobě následujících (nezávislých) sázkách je tedy $(19/37)^8$. Pravděpodobnost, že v těchto osmi hrách vyhraje 10 Kč (při daném postupu) je tedy $1-(19/37)^8$. Na to, aby vyhrál 2550 Kč, potřebuje 255 krát vyhrát po desetikoruně. Tedy opět podle pravidla součinu je pravděpodobnost výhry

$$\left(1-\left(\frac{19}{37}\right)^8\right)^{255} \doteq 0,29.$$

Tedy pravděpodobnost výhry je nižší, než kdyby vsadil rovnou vše na jednu barvu. □

9.13. Samostatně si můžete vyzkoušet spočítat předchozí příklad za předpokladu, že Aleš sází stejnou metodou jako v předchozím příkladě, končí však až v okamžiku, kdy nemá žádné peníze (pokud nemá na vsazení dvojnásobku částky prohrané v předchozí sázce, ale má ještě nějaké peníze, začíná sázet znovu od 10 Kč).

S podmíněnou pravděpodobností jsme se setkali již v první kapitole, viz 1.20.

9.14. Z definičního vztahu podmíněné pravděpodobnosti (viz 9.14) odvoďte pro jev A a jev B, který je disjunktním sjednocením jevů $B_1, B_2, \ldots, B_n$ vztah

$$P(A|B) = \sum_{i=1}^{n} P(A|B_i)P(B_i|B) \tag{9.1}$$

jevů A_i vztah

$$\begin{aligned} P(\cup_{i=1}^k A_i) = &\sum_{i=1}^{k} P(A_i) - \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} P(A_i\cap A_j) + \\ &+ \sum_{i=1}^{k-2}\sum_{j=i+1}^{k-1}\sum_{\ell=j+1}^{k} P(A_i\cap A_j\cap A_\ell) - \\ &- \cdots + \\ &+ (-1)^{k-1} P(A_1\cap A_2\cap\cdots\cap A_k). \end{aligned}$$

Stejně zůstává beze změny definice *stochasticky nezávislých jevů*, která vystihuje představu, že u nezávisle probíhajících jevů se jejich pravděpodobnosti mají násobit.

STOCHASTICKÁ NEZÁVISLOST

Jevy A a B jsou stochasticky nezávislé, jestliže platí

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

Je samozřejmé, že jev jistý a jev nemožný jsou stochasticky nezávislé na jakémkoliv jiném jevu.

Připomeňme, že výměnou jednoho z jevů A_i v systému po dvou stochasticky nezávislých jevů $A_1, A_2, \ldots$, za jev opačný A_i^c dostaneme opět systém stochasticky nezávislých jevů, a platí vztah (1.12) ze strany 23

$$\begin{aligned} P(A_1\cup\cdots\cup A_k) &= 1 - P(A_1^c\cap\cdots\cap A_k^c) = \\ &= 1-(1-P(A_1))\ldots(1-P(A_k)). \end{aligned}$$

Základním příkladem pro nás i nadále zůstává tzv. klasická konečná pravděpodobnost, kterou jsme se při tvorbě matematického modelu inspirovali. Připomeňme, že v tomto případě je Ω konečná množina, jevovým polem $\mathcal{A}$ je systém všech podmnožin v Ω a *klasická pravděpodobnost* je pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$ s pravděpodobnostní funkcí $P:\mathcal{A}\to\mathbb{R}$,

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

To odpovídá představě o relativní četnosti p_A jevu A při náhodném výběru prvků z množinu Ω.

Naše definice pravděpodobnosti zajišťuje rozumné chování na rostoucích či klesajících spočetných řetězcích jevů:

9.13. Věta. *Uvažme pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$ a neklesající řetězec jevů $A_1\subset A_2\subset\ldots$. Pak platí*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i\to\infty} P(A_i).$$

Pokud je naopak $A_1\supset A_2\supset A_3\supset\ldots$, potom platí

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

DŮKAZ. Přepíšeme uvažované sjednocení jevů $A=\cup_{i=1}^{\infty}A_i$ pomocí neslučitelných jevů

$$\tilde{A}_i = A_i \setminus A_{i-1},$$

definovaných pro všechna $i=2,3,\ldots$, a klademe $\tilde{A}_1 = A_1$. Potom

$$P(A) = P\left(\bigcup_{i=1}^{\infty}\tilde{A}_i\right) = \sum_{i=1}^{\infty} P(\tilde{A}_i) = \lim_{k\to\infty}\sum_{i=1}^{k} P(\tilde{A}_i).$$

Řešení. Všimněme si nejprve, že jevy $A \cap B_1$, $A \cap B_2$,…, $A \cap B_n$ jsou rovněž disjunktní. Můžeme tedy psát

$$\begin{aligned} P(A|B_1 \cup \cdots \cup B_n) &= \frac{P\left(A \cap (B_1 \cup \cdots \cup B_n)\right)}{P(B_1 \cup \cdots \cup B_n)} = \\ &= \frac{P\left((A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_n)\right)}{P(B)} = \\ &= \frac{\sum_{i=1}^{n} P(A \cap B_i)}{P(B_i)} \cdot \frac{P(B_i)}{P(B)} = \\ &= \sum_{i=1}^{n} P(A|B_i) P(B_i|B). \end{aligned}$$

□

9.15. Máme čtyři sáčky a v nich následující počty koulí: v prvním čtyři bílé, ve druhém tři bílé a jednu černou, ve třetím dvě bílé a dvě černé a ve čtvrtém čtyři černé. Náhodně vybereme sáček a z něj začneme bez vracení vytahovat koule. Určete pravděpodobnost, že

a) první dvě vytažené koule budou různých barev
b) a že druhá vytažená koule bude bílá, jestliže první vytažená koule byla bílá.

Řešení. Protože ve všech sáčcích je stejný počet koulí, je pravděpodobnost vytažení libovolné z koulí, potažmo libovolné dvojice koulí, stejná. Budeme tedy příklad řešit pomocí klasické pravděpodobnosti

a) Celkem můžeme vytáhnout 24 různých dvojic koulí, z toho je sedm dvojic složených z různobarevných koulí, hledaná pravděpodobnost je tedy 7/24.
b) Označme A jev, že první vytažená koule byla bílá, B jev, že druhá vytažená koule bude bílá. Potom $P(B \cap A)$ je pravděpodobnost, že první dvě vytažené koule budou bílé a ta je podobně jako v předchozím případě $10/24 = 5/12$. A opět klasickou pravděpodobností můžeme spočítat i $P(A)$, všech koulí je 16, z toho 9 bílých. Celkem

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{\frac{5}{12}}{\frac{9}{16}} = \frac{20}{27}.$$

Jiné řešení. Jev A můžeme uvážit jako sjednocení tří disjunktních jevů A_1, A_2, resp. A_3 a to, že jsme zvolili první sáček a z něj vytáhli bílou kouli, že jsme zvolili druhý sáček a z něj vytáhli bílou kouli a konečně že jsme zvolili třetí sáček a z něj vytáhli bílou kouli. Protože v každém sáčku je stejný počet koulí, je pravděpodobnost vytažení libovolné (bílé) koule shodná a tudíž $P(A) = \frac{9}{16}$ a

Přitom ale pro konečné součty máme

$$\sum_{i=1}^{\infty} P(\tilde{A}_i) = P(A_1) + \sum_{i=2}^{k} \big(P(A_i) - P(A_{i-1})\big) = P(A_n)$$

díky předpokládaným vztahům $A_{i-1} \subset A_i$. Tím jsme dokázali první tvrzení věty.

Ve druhém tvrzení můžeme přejít od jevů A_i k jejich komplementům $B_i = A_i^c$. Ty pak splňují předpoklady první části věty. Komplement k uvažovanému průniku je

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

Druhé tvrzení nyní plyne ze vztahu

$$P(A) = 1 - P(B) = \lim_{i \to \infty} \big(1 - P(B_i)\big) = 1 - \lim_{i \to \infty} P(B_i)$$

a důkaz je ukončen. □

9.14. Podmíněná pravděpodobnost. Popřemýšlejme nad následujícím úkolem. V předmětu X obvykle uspěje u zkoušky 40% studentů, v předmětu Y obvykle uspěje 80% studentů. Zaslechneme-li na chodbě jednoho ze studentů obou předmětů říkat, že u zkoušky uspěl, s jakou pravděpodobností šlo o předmět X?

Jak jsme stručně zmínili už v odstavci 1.20 na straně 23, umíme takové úlohy formalizovat následovně.

Definice. Nechť H je jev s nenulovou pravděpodobností v jevovém poli $\mathcal{A}$ v pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$. *Podmíněná pravděpodobnost* $P(A|H)$ jevu $A \in \mathcal{A}$ vzhledem k hypotéze H je definována vztahem

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

Definice odpovídá představě z klasické pravděpodobnosti, že jevy A a H nastanou zároveň, za předpokladu, že jev H nastal, s pravděpodobností $P(A \cap H)/P(H)$.

Je také vidět přímo z definice, hypotéza H a jev A jsou nezávislé tehdy a jen tehdy, je-li $P(A) = P(A|H)$.

Na první pohled se může zdát, že zavedením podmíněné pravděpodobnosti jsme nic nového nepřinesli. Ve skutečnosti jde o velice důležitý přístup, ke kterému se budeme vracet i ve statistice. Hypotéza totiž může mít charakter tzv. apriorní (tj. předem předpokládané) pravděpodobnosti a výsledné pravděpodobnosti pak říkáme aposteriorní (tj. bereme ji jako důsledek našeho předpokladu).

Přímo z definice vyplývá následující výsledek.

Lemma. *Nechť jev B je disjunktním sjednocením jevů B_1, B_2,…, B_n. Potom*

$$P(A|B) = \sum_{i=1}^{n} P(A|B_i) P(B_i|B) \tag{9.2}$$

$P(A_1|A) = \frac{\frac{4}{16}}{\frac{9}{16}} = \frac{4}{9}$, $P(A_2|A) = \frac{3}{9} = \frac{1}{3}$, $P(A_3|A) = \frac{2}{9}$. Použitím vztahu (9.2) pak dostáváme

$$\begin{aligned}
&P(B|A) = \\
&= P(B|A_1)P(A_1|A) + P(B|A_2)P(A_2|A) + P(B|A_3)P(A_3|A) = \\
&= P(B|A_1) \cdot \frac{P(A_1)}{P(A)} + P(B|A_2) \cdot \frac{P(A_2)}{P(A)} + P(B|A_3) \cdot \frac{P(A_3)}{P(P(A)} = \\
&= 1 \cdot \frac{4}{9} + \frac{2}{3} \cdot \frac{3}{9} + \frac{1}{3}\frac{2}{9} = \frac{20}{27}.
\end{aligned}$$

□

9.16. Mirek má čtyři sáčky, v každém jsou bílé a černé kuličky a to v těchto počtech: čtyři bílé; tři bílé a jedna černá; dvě bílé a dvě černé; jedna bílá a tři černé. Mirek náhodně jeden sáček vybral a náhodně z něj vytáhl jednu kouli. Byla černá. Mirek tento sáček zahodil a náhodně vybral jeden ze zbylých tří sáčků a z něj náhodně jednu kouli. Jaká je pravděpodobnost, že bude bílá?

Řešení. Podobně jako v předchozím příkladě, označíme jako A jev, že Mirek náhodně vybral sáček a z něj náhodně černou kouli. Tento jev disjunktním sjednocením jevů A_i, $i = 2, 3, 4$, kde A_i je jev, že Mirek vybral i-tý sáček a z něj potom černou kouli. Opět je pravděpodobnost vytažení libovolné (černé) koule stejná a tedy $P(A_2|A) = \frac{1}{6}$, $P(A_3|A) = \frac{2}{6} = \frac{1}{3}$ a $P(A_4|A) = \frac{3}{6} = \frac{1}{2}$. Nechť B je jev, že Mirek po zahození jednoho ze sáčků vybral ze zbylých bílou kouli. Pokud vyhodil druhý sáček, tak ve zbylých sáčcích je dohromady 7 bílých koulí a pravděpodobnost, že vytáhne jednu z nich je $P(B|A_2) = \frac{7}{12}$ (opět můžeme použít klasickou pravděpodobnost, protože v každém sáčku je stejný počet koulí a tedy má každá stejnou pravděpodobnost, že bude vytažena). Obdobně $P(B|A_3) = \frac{8}{12}$ a $P(B|A_4) = \frac{9}{12}$. Pak podle (5) je hledaná pravděpodobnost

$$\begin{aligned}
&P(B|A) = \\
&= P(B|A_2)P(A_2|A) + P(B|A_3)P(A_3|A) + P(B|A_4)P(A_4|A) = \\
&= \tfrac{7}{12} \cdot \tfrac{1}{6} + \tfrac{8}{12} \cdot \tfrac{1}{3} + \tfrac{9}{12} \cdot \tfrac{1}{2} = \tfrac{25}{36}.
\end{aligned}$$

□

9.17. Mirek má čtyři sáčky, v každém jsou bílé a černé kuličky a to v těchto počtech: jedna bílá a jedna černá; tři bílé a jedna černá; jedna bílá a dvě černé; jedna bílá a tři černé. Mirek náhodně jeden sáček vybral a náhodně z něj vytáhl jednu kouli. Byla bílá. Mirek tento sáček zahodil a náhodně vybral jeden ze zbylých tří sáčků a z něj náhodně jednu kouli. Jaká je pravděpodobnost, že bude bílá?

Řešení. Podobně jako v předchozím příkladě uvážíme jev A, totiž že Mirek vybral náhodně sáček a z něj náhodně bílou kouli jako sjednocení čtyř disjunktních jevů A_1, A_2, A_3 a A_4: Mirek vytáhl bílou kouli a před tím zahodil první, resp. druhý, resp. třetí, resp. čtvrtý sáček.

Důkaz. Všimněme si nejprve, že jevy $A \cap B_1, A \cap B_2, \ldots, A \cap B_n$ jsou rovněž disjunktní. Můžeme tedy psát

$$\begin{aligned}
P(A|B_1 \cup \cdots \cup B_n) &= \frac{P\left(A \cap (B_1 \cup \cdots \cup B_n)\right)}{P(B_1 \cup \cdots \cup B_n)} = \\
&= \frac{P\left((A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_n)\right)}{P(B)} = \\
&= \frac{\sum_{i=1}^{n} P(A \cap B_i)}{P(B_i)} \cdot \frac{P(B_i)}{P(B)} = \\
&= \sum_{i=1}^{n} P(A|B_i)P(B_i|B).
\end{aligned}$$

□

Uvažujme zvláštní případ $B = \Omega$. Pak jevy B_i můžeme interpretovat jako „možné stavy světa“, $P(A|B_i)$ vyjadřuje pravděpodobnost jevu A, pokud je svět v i–tém stavu, $P(B_i|\Omega) = P(B_i)$ je pravděpodobnost toho, že svět se v i–tém stavu nachází. Podle předchozího lemmatu platí

$$P(A) = P(A|\Omega) = \sum_{i=1}^{n} P(A|B_i)P(B_i).$$

Tento vztah se nazývá vzorec pro celkovou pravděpodobnost (nebo věta o úplné pravděpodobnosti).

9.15. Bayesova věta. Jednoduchým přepsáním vzorce pro podmíněnou pravděpodobnost dostáváme

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

Odtud okamžitě plyne velice důležitý důsledek:

Bayesův vzorec

Věta. *Pro pravděpodobnost jevů A a B platí*

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{9.3}$$

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^c)P(B|A^c)}. \tag{9.4}$$

Prvnímu tvrzení se také říká vzorec pro *inverzní pravděpodobnosti*, zatímco druhé tvrzení je označováno jako 1. *Bayesův vzorec*.

Důkaz. První tvrzení je jen přepsáním výpočtu před větou. Abychom dostali druhé tvrzení všimněme si, že

$$P(B) = P(B \cap A) + P(B \cap A^c),$$

proto můžeme podle vzorce pro celkovou pravděpodobnost dosadit $P(B) = P(A)P(B|A) + P(A^c)P(B|A^c)$ do vzorce pro inverzní pravděpodobnost a dostáváme právě druhé tvrzení věty. □

Bayesův vzorec bývá často formulován v lehce obecnějším tvaru, který se dokáže stejným způsobem jako (9.4):

Nechť je základní prostor Ω sjednocením disjunktních jevů $A_1, \ldots A_n$. Pak pro libovolné $i \in \{1, \ldots, n\}$ platí

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}. \tag{9.5}$$

Pravděpodobnost vytažení bílé koule z prvního sáčku je $P(A_1) = \frac{1}{4}\cdot\frac{1}{2}$ (jev A_1 je dán tím, že současně nastaly dva nezávislé jevy a to, že vytáhl první sáček a že z prvního sáčku vytáhl bílou kouli), podobně $P(A_2) = \frac{1}{4}\cdot\frac{3}{4}$, $P(A_3) = \frac{1}{4}\cdot\frac{1}{3}$, $P(A_4) = \frac{1}{4}\cdot\frac{1}{4}$. $P(A) = P(A_1)+P(A_2)+P(A_3)+$ $+P(A_4) = \frac{11}{24}$. Všimněme si, že pravděpodobnost $P(A)$ nemůžeme počítat klasickou pravděpodobností, tedy prostým podělením počtu bílých koulí ku počtu všech koulí, protože například pravděpodobnost vytažení dané koule v prvním sáčku je dvojnásobná oproti vytažení dané koule ze čtvrtého sáčku. Pro podmíněné pravděpodobnosti pak platí $P(A_1|A) = P(A_1)/P(A) = \frac{3}{11}$, $P(A_2|A) = \frac{9}{22}$, $P(A_3|A) = \frac{2}{11}$, $P(A_4|A) = \frac{3}{22}$. Označíme ještě písmenem B jev, že Mirek po zahození jednoho ze sáčků vytáhne bílou kouli a znovu budeme chtít použít vztah (5). Zbývá ještě dopočítat $P(B|A_i)$, $i = 1, \ldots 4$. Jev $P(B|A_1)$ rozdělíme na tři disjunktní jevy B_2, B_3, B_4, totiž že druhá vytažená koule byla z druhého, resp. třetího, resp. čtvrtého sáčku. Celkem

$$P(B|A_1) = P(B_2|A_1) + P(B_3|A_1) + P(B_4|A_1) = \\ = \frac{1}{3}\cdot\frac{3}{4} + \frac{1}{3}\cdot\frac{1}{3} + \frac{1}{3}\cdot\frac{1}{4} = \frac{4}{9}.$$

Obdobně

$$P(B|A_2) = \frac{1}{3}\cdot\frac{1}{2} + \frac{1}{3}\cdot\frac{1}{3} + \frac{1}{3}\cdot\frac{1}{4} = \frac{13}{36},$$
$$P(B|A_3) = \frac{1}{3}\cdot\frac{1}{2} + \frac{1}{3}\cdot\frac{3}{4} + \frac{1}{3}\cdot\frac{1}{4} = \frac{1}{2},$$
$$P(B|A_4) = \frac{1}{3}\cdot\frac{1}{2} + \frac{1}{3}\cdot\frac{3}{4} + \frac{1}{3}\cdot\frac{1}{3} = \frac{19}{36}.$$

Celkem pak

$$P(B|A) = P(B|A_1)P(A_1|A) + P(B|A_2)P(A_2|A) + \\ + P(B|A_3)P(A_3|A) + P(B|A_4)P(A_4|A) = \\ = \frac{4}{9}\cdot\frac{3}{11} + \frac{13}{36}\cdot\frac{9}{22} + \frac{1}{2}\cdot\frac{2}{11} + \frac{19}{36}\cdot\frac{3}{22} = \frac{19}{44}.$$

□

9.18. Dva střelci vystřelí každý dvě rány na terč. První má pravděpodobnost zásahu 80%, druhý 60%. V terči se našly dvě rány. Jaká je pravděpodobnost, že obě patří prvnímu střelci?

Řešení. Pravděpodobnost zásahu prvního střelce jsou tedy 4/5, druhého 3/5. Uvažme dva jevy:

A ... v terči se našly dva zásahy patřící prvnímu střelci,

B ... v terči se našly dva zásahy.

Dle zadání úlohy máme zjistit $P(A|B)$. Rozdělme jev B na šest disjunktních jevů podle toho, který střelec a který svůj výstřel do terče umístil. Jevy uvedeme v tabulce a u každého navíc spočítáme pravděpodobnost toho, že nastane. Uvědomíme si při tom, že každá

9.16. Poznámky. Nyní se můžeme snadno vypořádat s úvodní

otázkou z minulého odstavce. Dotaz si nejprve malinko upřesníme. Uvažujeme jev A představující „student u zkoušky uspěl“ a jev B, který říká „student byl zkoušen z předmětu X“. Předpokládáme přitom, že pravděpodobnosti zkoušení z obou předmětů jsou stejné, tj. $P(B) = P(B^c) = 0{,}5$. Zatímco hledaná pravděpodobnost $P(B|A)$ je zatím spíše nejasná, pravděpodobnost $P(A|B) = 0{,}4$ je dána přímo v zadání.

To je typický případ použití Bayesových vzorců. Když přitom použijeme druhý z nich, vůbec nemusíme počítat $P(A)$:

$$P(B|A) = \frac{P(B)P(A|B)}{P(B)P(A|B) + P(B^c)P(A|B^c)} = \\ = \frac{0{,}5\cdot 0{,}4}{0{,}5\cdot 0{,}4 + 0{,}5\cdot 0{,}8} = \frac{1}{3}.$$

Abychom si přiblížili roli apriorní pravděpodobnosti hypotézy, podívejme se ještě na jeden příklad.

Řekněme, že testy připravenosti a znalostí, na základě kterých jsou studenti přijímáni na univerzitu, mají následující spolehlivost v testování inteligence osob: 99% inteligentních osob má pozitivní výsledek testu, zatímco u neinteligentních uchazečů má 0,5% z nich pozitivní výsledek testu. Chceme zjistit, s jakou pravděpodobností je náhodně vybraný student na univerzitě inteligentní.

Máme tedy jev A „náhodně zvolená osoba je inteligentní“ a jev B „osoba prošla testem s pozitivním výsledkem“. Dle Bayesova vzorce můžeme opět rovnou spočíst pravděpodobnost, že nastal jev A za předpokladu, že nastal jev B. Musíme jen dodat všeobecnou pravděpodobnost $p = p(A)$, že náhodně zvolený uchazeč o studium je inteligentní.

$$P(A|B) = \frac{p\cdot 0{,}99}{p\cdot 0{,}99 + (1-p)\cdot 0{,}005}.$$

V následující tabulce je spočten pro různé hodnoty p vyjádřené v jednotkách promile. V prvním sloupci tedy je výsledek za předpokladu, že je mezi uchazeči o studium každý druhý inteligentní atd.

p	500	100	50	10	1	0,1
$P(A\|B)$	0,99	0,96	0,91	0,67	0,17	0,02

Pokud tedy je každý druhý uchazeč inteligentní, máme na univerzitě používající náš test 99% inteligentních studentů. Pokud ale naší představě o inteligenci odpovídá jen 1% populace a uchazeči jsou dobrým náhodným vzorkem, pak už máme na univerzitě jen zhruba dvě třetiny inteligentních studentů ...

Představme si ale, že obdobné testování provedeme při plošném testování výskytu nějaké nemoci, třeba HIV. Dejme tomu, že máme stejně citlivý test jako výše a prověříme jím o přestávce mezi přednáškami všechny přítomné studenty. V tomto případě bychom měli předpokládat, že parametr p bude obdobný jako u celé populace, tj. řekněme jeden nakažený z 10000 obyvatel, což odpovídá poslednímu sloupci v tabulce. Pak ovšem je výsledek testu katastroficky nespolehlivý. Jen asi u 2 procent pozitivně otestovaných se jedná o skutečně nemocné studenty!

Všimněme si, že problém je zapříčiněn jakýmkoliv malým výskytem pozitivních výsledků u zdravých osob. I kdybychom zlepšili test tak, že bude na 100% účinný při testu pozitivní osoby, neovlivníme skoro vůbec výsledné pravděpodobnosti v tabulce.

Při lékařské diagnostice vzácných chorob je při pozitivním výsledku testu nutné provést další test. Přitom výsledek prvního testu

uvažovaná střelba se skládá ze čtyř nezávislých jevů: výsledek střelby hráče A či B v prvním či druhém výstřelu. V tabulce značíme zásah jedničkou, minutí terče nulou.

	1. střelec		2.střelec		pst nastoupení jevu
B_1	0	1	0	1	$\frac{1}{5}\cdot\frac{4}{5}\cdot\frac{2}{5}\cdot\frac{3}{5}$
B_2	0	1	1	0	$\frac{24}{25^2}$
B_3	1	0	1	0	$\frac{24}{25^2}$
B_4	1	0	0	1	$\frac{24}{25^2}$
B_5	1	1	0	0	$\frac{64}{25^2}$
B_6	0	0	1	1	$\frac{9}{25^2}$

Sečtením pravděpodobnosti těchto disjunktních jevů dostáváme:

$$P(B)=\sum_{i=1}^{6}P(B_i)=169/625.$$

Nyní můžeme přistoupit k výpočtu podmíněné pravděpodobnosti podle vztahu z odstavce 9.14:

$$P(A|B)=\frac{P(A\cap B)}{P(B)}=\frac{P(B_5)}{P(B)}=\frac{\frac{64}{625}}{\frac{169}{625}}=\frac{64}{169}\doteq 0{,}38.$$

□

9.19. Hodíme mincí. Pokud padne líc, dáme do krabice bílou kulečníkovou kouli, pokud padne rub, dáme tam kouli černou. To opakujeme n-krát. Potom poslepu vybereme z krabice jednu kouli a nevrátíme ji zpět. Tato vybraná koule je bílá. Určete pravděpodobnost, že další poslepu vybraná koule je černá.

Řešení. V zadání není řečeno, o jakou minci jde. Aby úloha vůbec měla nějaký rozumný smysl,budeme předpokládat, že výsledky hodů touto mincí jsou nezávislé a že existuje pravděpodobnost padnutí lícové strany. Tuto pravděpodobnost označíme p. Ze zadaného faktu, že první vytažená koule je bílá, usoudíme, že $p>0$. Poznámku o tom, že koule jsou kulečníkové, budeme chápat tak, že jednotlivé koule jsou hmatem nerozlišitelné a tedy vyjádření „vybereme poslepu“ označuje totéž, co „vybereme náhodně“. Jev „koule v krabici je bílá“ odpovídá jevu „v příslušném hodu mincí padl líc“. To znamená, že pravděpodobnostní prostor „náhodné vytažení koule z krabice“ je izomorfní pravděpodobnostnímu prostoru „hod mincí“. Z předpokládané nezávislosti výsledků jednotlivých hodů mincí plyne nezávislost barev tažených koulí. Touto úvahou dostáváme, že hledaná pravděpodobnost je rovna $1-p$.

Je provedená úvaha přesvědčivá? Očekáváme přece, že v plné krabici je np bílých a $n(1-p)$ černých koulí (přesněji řečeno, celé části těchto hodnot). Jednu bílou kouli jsme odstranili, takže relativní $P(A|B)$ má roli apriorní pravděpodobnosti $P(A)$ při druhém testu. Tento postup umožňuje „kumulovat zkušenost“.

V obou případech tedy musíme při přípravě testu dbát na to, abychom si zajistili přiměřeně vysoké p. U procesu přijímání studentů na univerzitu to asi bude dobrý marketing univerzity, který zajistí, aby se neinteligentní osoby hlásily v daleko menší míře, než je jejich výskyt v populaci. U testování chorob nejspíš půjde o souběh dalších skutečností a činností (např. testování HIV pozitivitu pouze u rizikových skupin obyvatelstva a podobně).

9.17. Borelovské množiny. Vraťme se k jednoduchému a názornému příkladu statistik kolem výsledků studentů v daném předmětu. Ten je a není podobný klasické pravděpodobnosti a s ní související statistice při házení kostkou.

Na jedné straně máme pouze konečný počet studentů a připustili jsme pouze konečný počet možných bodových hodnocení práce studenta za semestr (např. celá čísla od 0 do 20). Zároveň ale není patrně vhodné představovat si výsledky jednotlivých studentů jako analogii nezávislého házení pravidelnou kostkou. Jednak neexistuje pravidelný 21–stěn, ale hlavně by to byla skutečně divně vedená přednáška.

Na základním (konečném) prostoru Ω všech studentů máme prostě definovánu funkci bodového ohodnocení $X:\Omega\to\mathbb{R}$, která má tu vlastnost, že můžeme modelovat pravděpodobnosti příslušnosti její hodnoty do předem zvoleného intervalu při náhodném výběru studenta. Např. můžeme chtít modelovat pravděpodobnost, že student uspěl s hodnocením A nebo B. Pokud známe výsledky všech studentů, snadno dostaneme statistiky celého souboru, např. výběrový průměr $\bar{X}$ a směrodatnou odchylku S_x.

Patrně bychom od rozumně vedené přednášky a dobrých studentů očekávali, že nejvyšší pravděpodobnost výsledku bude ležet někde uprostřed škály v „úspěšném intervalu“, zatímco ideální výsledek plného bodového zisku příliš pravděpodobný nebude. Stejně tak bude příliš mnoho hodnot X v intervalu neúspěšných hodnot na většině univerzit bráno jako výrazný neúspěch přednášejícího.

Často ale v podobných situacích máme k dispozici jen náhodně vybraných několik studentů a známe příslušné statistiky jen pro tento vybraný vzorek. Pak se můžeme dívat na příslušné hodnoty jako na vektor $(X_1,\dots,X_k)$ a bude nás opět zajímat jakákoliv funkce na tomto vektoru (např. některá z výše zmíněných statistik).

Je to typický příklad tzv. náhodných veličin a náhodných vektorů, jak je budeme definovat v dalším odstavci. Budeme chtít umět diskutovat pravděpodobnost, že hodnota X padne do kteréhokoliv intervalu $(a,b)\subset\mathbb{R}$ s reálnými čísly a,b a uzavřenými nebo otevřenými konci intervalu. Případně budeme potřebovat totéž pro vícerozměrné intervaly v $\mathbb{R}^k$ a vektory $(X_1,\dots,X_k)$.

Zkusme tedy uvažovat číselné veličiny X na nějakém základním prostoru, tj. obyčejné funkce $X:\Omega\to\mathbb{R}$. Chceme pracovat s pravděpodobnostmi příslušnosti hodnoty X do předem zadaného intervalu. Musíme proto uvést do souladu požadavky na pravděpodobnostní prostor všech jevových polí s vlastnostmi takových funkcí:

Borelovské množiny v $\mathbb{R}^k$

Na prostoru $\mathbb{R}^k$ uvažujme nejmenší jevové pole $\mathcal{B}$ obsahující všechny k–rozměrné intervaly. Množinám v $\mathcal{B}$ říkáme *Borelovské množiny* na $\mathbb{R}^k$. Speciálně pro $k=1$ půjde o všechny množiny,

četnost černých koulí o něco vzroste a proto i pravděpodobnost vytažení černé koule bude větší než $1-p$. Než budete číst dále, pokuste se uhodnout, zda pravděpodobnost vytažení černé koule bude $1-p$ nebo větší, případně jak tuto pravděpodobnost ovlivní hodnota n (počet koulí v krabici před vytahováním).

Úlohu budeme nyní řešit poněkud sofistikovaněji. Označme B_i jev „v plné krabici je i bílých koulí“ (zřejmě $i \in \{0, 1, 2, \ldots, n\}$), A jev „první vytažená koule je bílá“ a C jev „druhá vytažená koule je černá“. Jev B_i je vlastně jevem, že v sérii n hodů mincí padl líc i-krát, tedy

$$P(B_i) = \binom{n}{i} p^i (1-p)^{n-i}.$$

Podmíněná pravděpodobnost vytažení bílé koule za podmínky, že v krabici je právě i bílých koulí, je rovna

$$P(A|B_i) = \frac{i}{n}.$$

Zajímá nás pravděpodobnost jevu C když víme, že nastal jev A, tedy $P(C|A)$. Poněvadž jevy B_i jsou neslučitelné, jsou neslučitelné i jevy $C \cap B_i$. Současně platí $C = \bigcup_{i=0}^{n} (C \cap B_i)$ a toto sjednocení je disjunktní. Proto můžeme psát

$$\begin{aligned} P(C|A) &= P\left(\bigcup_{i=0}^{n} (C \cap B_i)|A\right) = \sum_{i=0}^{n} \frac{P\big((C \cap B_i) \cap A\big)}{P(A)} = \\ &= \frac{1}{P(A)} \sum_{i=0}^{n} P\big(C \cap (A \cap B_i)\big) = \\ &= \frac{1}{P(A)} \sum_{i=0}^{n} P(A \cap B_i) P(C|A \cap B_i) = \\ &= \frac{1}{P(A)} \sum_{i=0}^{n} P(B_i) P(A|B_i) P(C|A \cap B_i). \end{aligned}$$

Za pravděpodobnost $P(A)$ můžeme ještě dosadit ze vzorce pro celkovou pravděpodobnost a dostaneme

$$\begin{aligned} P(C|A) &= \frac{\sum_{i=0}^{n} P(B_i) P(A|B_i) P(C|A \cap B_i)}{P(A)} = \\ &= \frac{\sum_{i=0}^{n} P(B_i) P(A|B_i) P(C|A \cap B_i)}{\sum_{i=0}^{n} P(B_i) P(A|B_i)}. \end{aligned} \tag{9.2}$$

Tato formulka bývá někdy nazývána 2. *Bayesův vzorec*; obecně platí za předpokladu, že prostor Ω je disjunktním sjednocením jevů B_i.

které ze všech intervalů obdržíme konečnými průniky a nejvýše spočetnými sjednoceními.[1]

9.18. Náhodné veličiny. Nyní už máme všechno připraveno pro definici náhodných veličin a náhodných vektorů. Poznamenejme již předem, že pro klasickou konečnou pravděpodobnost je náhodnou veličinou každá reálná funkce $X : \Omega \to \mathbb{R}$. Skutečně, na konečné množině Ω nabývá X jen konečně mnoho hodnot a každá podmnožina v Ω je jevem.

NÁHODNÉ VELIČINY A VEKTORY

Definice. *Náhodná veličina* X na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$ je taková funkce $X : \Omega \to \mathbb{R}$, že vzor $X^{-1}(B)$ patří do $\mathcal{A}$ pro každou Borelovskou množinu $B \in \mathcal{B}$ na $\mathbb{R}$. Reálná funkce $P_X(B) = P(X^{-1}(B))$ definovaná na všech intervalech $B \subset \mathbb{R}$ se nazývá *rozdělení (pravděpodobnosti) náhodné veličiny* X.

Náhodný vektor $X = (X_1, \ldots, X_k)$ na $(\Omega, \mathcal{A}, P)$ je k–tice náhodných veličin $X_i : \Omega \to \mathbb{R}$ definovaných na stejném základním pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$.

Jestliže vybereme intervaly $I_1, \ldots, I_k$ v $\mathbb{R}$ a definujeme množinu $B = I_1 \times \cdots \times I_k$, pak jistě existuje pravděpodobnost současného výskytu všech k jevů $X_i \in I_i$. Díky aditivitě funkce P tedy bude, obdobně jako ve skalárním případě, existovat reálná funkce $P_X(B) = P(X^{-1}(B))$ definovaná na všech k–rozměrných intervalech $B \subset \mathbb{R}^k$. Nazýváme ji *rozdělení (pravděpodobnosti) náhodného vektoru* X.

9.19. Distribuční funkce. Rozdělení náhodných veličin zadáváme nejčastěji pomocí pravidla, jak roste pravděpodobnost s přírůstkem intervalu B.

Definice náhodné veličiny zajišťuje, že pro všechny intervaly I s krajními body a, b, $-\infty \le a \le b \le \infty$, existuje pravděpodobnost jevu $P(I)$. Budeme ji zapisovat stručně $P(a < X < b)$, resp. $P(X < b)$ pokud je $a = -\infty$, pro otevřený interval I a obdobně pro intervaly uzavřené nebo z jedné strany uzavřené. Ve speciálním případě jediné hodnoty píšeme $P(X = a)$.

Podobně u náhodného vektoru $X = (X_1, \ldots, X_k)$ píšeme stručně $P(a_1 < X_1 < b_1, \ldots, a_k < X_k < b_k)$, pro současné nastoupení jevů, kdy hodnoty X_i padnou do uvedených intervalů (které mohou být také uzavřené neohraničené apod.).

DISTRIBUČNÍ FUNKCE

Definice. *Distribuční funkcí* náhodné veličiny X je funkce F_X $\mathbb{R} \to [0, 1]$ definovaná pro všechny $x \in \mathbb{R}$ vztahem[2]

$$F_X(x) = P(X < x).$$

Distribuční funkcí náhodného vektoru $(X_1, \ldots, X_k)$ je funkce $F_X : \mathbb{R}^k \to \mathbb{R}$ definovaná pro všechny $x = (x_1, \ldots, x_k) \in \mathbb{R}^k$ vztahem

$$F_X(x) = P(X_1 < x_1, \ldots, X_k < x_k).$$

[1] V této souvislosti se často také hovoří o tzv. *σ–algebře Borelovsky měřitelných množin na $\mathbb{R}^k$* a následující definici lze formulovat tak, že náhodné veličiny jsou Borelovsky měřitelné funkce.

[2] V literatuře se stejně často potkáváme také s definicí s neostrou nerovností, tj. pravděpodobnost $P(X = x)$ je ještě započtena také. V takovém případě platí obdobné vlastnosti jako ve větě 9.20, jen je distribuční funkce zprava spojitá apod.

Ještě si uvědomíme, že podle zadání úlohy jsme alespoň jednou hodili mincí a tedy $n \geq 1$. Nyní můžeme vypočítat

$$\begin{aligned}\sum_{i=0}^{n} P(B_i)P(A|B_i) &= \sum_{i=0}^{n} \binom{n}{i} p^i(1-p)^{n-i} \cdot \frac{i}{n} = \\ &= \sum_{i=1}^{n} \frac{(n-1)!}{(i-1)!(n-i)!} p^i(1-p)^{n-i} = \\ &= \sum_{i=0}^{n-1} \frac{(n-1)!}{i!(n-i-1)!} p^{i+1}(1-p)^{n-i-1} = \\ &= p\sum_{i=0}^{n-1} \binom{n-1}{i} p^i(1-p)^{n-1-i} = \\ &= p\big(p+(1-p)\big)^{n-1} = p,\end{aligned}$$

$$\begin{aligned}&\sum_{i=0}^{n} P(B_i)P(A|B_i)P(C|A\cap B_i) = \\ &= \sum_{i=0}^{n} \binom{n}{i} p^i(1-p)^{n-i} \cdot \frac{i}{n} \cdot \frac{n-i}{n-1} = \\ &= \sum_{i=1}^{n-1} \frac{(n-2)!}{(i-1)!(n-i-1)!} p^i(1-p)^{n-i} = \\ &= \sum_{i=0}^{n-2} \frac{(n-2)!}{i!(n-2-i)!} p^{i+1}(1-p)^{n-i-1} = \\ &= p(1-p)\sum_{i=0}^{n-2} \binom{n-2}{i} p^i(1-p)^{n-2-i} = \\ &= \begin{cases} p(1-p), & n > 1 \\ 0, & n = 1, \end{cases}\end{aligned}$$

takže po dosazení do druhého Bayesova vzorce dostaneme hledanou pravděpodobnost

$$P(C|A) = \begin{cases} 0, & n = 1, \\ 1-p, & n > 1. \end{cases}$$

Jednoduchá úvaha o izomorfii pravděpodobnostních prostorů tedy dala správný výsledek; výpočet pouze upozornil na triviální případ $n = 1$. □

9.20. V jedné vědomostní soutěži bylo hlavní výhrou Ferrari 599 GTB Fiorano. Soutěžící, který se dostal do posledního kola, byl přiveden před tři stejná vrata. Podmínkou získání výhry bylo správně uhodnout, za kterými vraty se automobil nachází. Soutěžící jedna vrata označil a poté asistent otevřel ta z neoznačených vrat, za nimiž byla koza. Poslední soutěžní otázkou bylo, zda soutěžící chce svůj tip měnit.

Je-li z kontextu zřejmé, které veličiny se distribuční funkce týká, její označení vypouštíme, tj. píšeme $F(x)$.

Další věta nás ujistí, že pro každou náhodnou veličinu umíme výhradně z distribuční funkce počítat pravděpodobnosti, že hodnoty X padnou do jakéhokoliv intervalu, tj. ve skutečnosti do jakékoliv Borelovské množiny B.

9.20. Věta. *Pro každou náhodnou veličinu X má její distribuční funkce $F : \mathbb{R} \to [0,1]$ následující vlastnosti*

(1) F je neklesající funkce;
(2) F má v každém bodě $x \in \mathbb{R}$ limitu zleva i limitu zprava;
(3) F je zleva spojitá;
(4) v nevlastních bodech má F limity

$$\lim_{x\to\infty} F(x) = 1, \qquad \lim_{x\to-\infty} F(x) = 0; \tag{9.6}$$

(5) pravděpodobnost, že X nabývá právě hodnotu x je dána

$$P(X = x) = \lim_{y\to x_+} F(y) - F(x). \tag{9.7}$$

(6) Distribuční funkce náhodné veličiny má vždy nejvýše spočetně mnoho bodů nespojitosti.

Důkaz. Důkaz spočívá ve vcelku jednoduchým přímých výpočtech. Zejména si uvědomme, že jevy $a \leq X < b$ a $X < a$ jsou disjunktní a proto platí

$$P(a \leq X < b) = P(X < b) - P(X < a) = F(b) - F(a).$$

Odtud již okamžitě z definice pravděpodobnosti plyne první dokazovaná vlastnost.

Další dvě tvrzení odvodíme z vlastností pravděpodobnosti na rostoucích či klesajících řetězcích jevů, které jsme odvodili ve větě 9.13. Zvolme nerostoucí posloupnost čísel $r_n > 0$ konvergující k 0 a uvažujme jevy A_n zadané požadavkem $X < x - r_n$. Sjednocení všech těchto jevů je právě jev A zadaný nerovností $X < x$. Jev A přitom pochopitelně nezávisí na volbě posloupnosti r_n. Podle prvního tvrzení věty 9.13 tedy bude

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

To však podle testu konvergence funkcí pomocí posloupností (viz str. 256) znamená, že limita zleva funkce F_X v bodě x existuje a je rovna $P(A)$. To dokazuje polovinu tvrzení (2) a zároveň tvrzení (3).

Zcela obdobně můžeme pomocí zvolené posloupnosti čísel r_n definovat jevy A_n odpovídající hodnotám $X_n < x + r_n$. tentokrát máme nerostoucí řetězec $A_1 \supset A_2 \supset \ldots$ a jejich průnikem bude jev $X \leq x$. Pro pravděpodobnost jevu A platí, podle druhé vlastnosti z věty 9.13,

$$P(A) = \lim_{n\to\infty} P(A_n) = P(X \leq x),$$

což ověřuje že limita zprava funkce F v bodě x existuje. Zároveň jsme přitom ověřili i vlastnost (5).

Limitní hodnoty z vlastnosti (4) věty se odvodí zcela obdobně s použitím věty 9.13, jak jsme výše spočetli limity zleva a zprava výše. V prvém případě půjde o jevy A_n zadané pomocí $X < r_n$, pro jakoukoliv rostoucí posloupnost $r_n \to \infty$. Jejich sjednocením bude jev jistý Ω. Ve druhém případě půjde o jevy A_n zadané pomocí $X < r_n$ pro jakoukoliv klesající posloupnost $r_n \to -\infty$ a jejich průnikem bude jev nemožný.

Řešení. Nevyslovený předpoklad úlohy je ten, že soutěžící zmíněný automobil chce získat. Nejdříve zkuste ověřit, jak spolehlivou intuici pro náhodné jevy již máte. Můžete uvažovat například takto: „Za jedněmi ze zbývajících dvou vrat je Ferrari, za každými z nich se stejnou pravděpodobností. Proto je jedno, která vrata jsou označená a nemá smysl svůj tip měnit.“ Nebo: „Pravděpodobnost, že jsem si hned na začátku tipnul správně je $\frac{1}{3}$. Na této pravděpodobnosti ta ukázaná koza nic nezmění, takže pravděpodobnost, že jsem tipoval špatně je $\frac{2}{3}$. Proto když tip změním, tak s pravděpodobností $\frac{2}{3}$ vyhraji.“

Změnit tip je rozumné pouze v případě, že pravděpodobnost automobilu za neoznačenými a neotevřenými vraty je větší, než jeho pravděpodobnost za vraty označenými. Pro výpočet si označíme jevy H „původní tip je správný“, A „tip byl změněn“ a C „soutěžící vyhrál“. Zajímají nás tedy pravděpodobnosti $P(C|A)$ a $P(C|A^c)$.

Soutěžící nejprve označil jedna vrata ze tří, Ferrari je jen za jedněmi z nich. Tedy

$$P(H) = \frac{1}{3}, \quad P(H^c) = 1 - \frac{1}{3} = \frac{2}{3}.$$

Změnu tipu považujeme za jev nezávislý na tipu původním, tedy

$$P(A|H) = P(A|H^c) = P(A), \quad P(A^c|H) = P(A^c|H^c) = P(A^c).$$

Pokud původní tip byl správný a soutěžící rozhodnutí změnil, pak nemohl vyhrát; naopak, pokud původní tip byl špatný a soutěžící rozhodnutí změnil, vyhrál jistě, tedy

$$P(C|A \cap H) = 0 = P(C|A^c \cap H^c),$$
$$P(C|A^c \cap H) = 1 = P(C|A \cap H^c).$$

Ze druhého Bayesova vzorce (‖9.2‖) nyní dostaneme

$$P(C|A) = \frac{P(H)P(A|H)P(C|A \cap H) + P(H^c)P(A|H^c)P(C|A \cap H^c)}{P(A)} = P(H^c) = \frac{2}{3}$$

a analogicky

$$P(C|A^c) = \frac{P(H)P(A^c|H)P(C|A^c \cap H) + P(H^c)P(A^c|H^c)P(C|A^c \cap H^c)}{P(A^c)} = P(H) = \frac{1}{3}.$$

Dostali jsme, že $P(C|A) > P(C|A^c)$, a proto je výhodné změnit tip. □

9.21. Máme dva sáčky. V jednom jsou dvě bílé a dvě černé v druhém jedna bílá a dvě černé. Náhodně vybereme sáček a z něj postupně (bez

Zbývá dokázat poslední tvrzení. Podle již dokázaných vlastností jsou body nespojitosti distribuční funkce právě ty hodnoty x, ve kterých má náhodná veličina tuto hodnotu s nenulovou pravděpodobností, tj. $P(X = x) \neq 0$. Označme nyní M_n množinu těch bodů x, pro které je $P(X = x) > \frac{1}{n}$. Evidentně je množina M všech bodů nespojitosti dána jako sjednocení $M = \cup_{n=2}^{\infty}$. Protože je ale součet pravděpodobností disjunktních jevů vždy nejvýše jedna, nemůže obsahovat M_n více než $n - 1$ prvků. Je tedy M spočetným sjednocením konečných množin a je tedy sama spočetná. □

Nyní je zřejmé, že můžeme z distribuční funkce snadno spočíst pravděpodobnost, že hodnota náhodné veličiny padne do jakéhokoliv daného intervalu. Zadává tedy skutečně distribuční funkce F_X celé rozložení pravděpodobnostní náhodné veličiny X.

9.21. Diskrétní a spojité náhodné veličiny. Náhodné veličiny se

chovají zásadně odlišně podle toho, jestli je veškerá nenulová pravděpodobnost „soustředěna do několika konečných hodnot“ nebo je naopak „spojitě rozprostřena“ po (části) reálné osy.

Diskrétní náhodné veličiny

Jestliže náhodná veličina X na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$ nabývá jen konečně mnoha různých hodnot $x_1, x_2, \ldots, x_n \in \mathbb{R}$ nebo případně spočetně mnoha reálných hodnot $x_1, x_2, \ldots$, říkáme, že jde o *diskrétní náhodnou veličinu*.

Definujeme pak *pravděpodobnostní funkci* $f(x)$ vztahem

$$f(x) = \begin{cases} P(X = x_i) & x = x_i \\ 0 & \text{jinak.} \end{cases}$$

Protože je pravděpodobnost spočetně aditivní a jednotlivé jevy $X = x_i$ jsou disjunktní, je součet všech hodnot $f(x_i)$ dán buď konečným součtem nebo absolutně konvergentní řadou

$$\sum_i f(x_i) = 1.$$

Pro rozdělení pravděpodobnosti veličiny X platí

$$P(X^{-1}(B)) = \sum_{x_i \in B} f(x_i)$$

a tedy zejména je distribuční funkce tvaru

$$F_X(t) = \sum_{x_i < t} f(x_i).$$

Všimněme si, že distribuční funkce $F(x)$ diskrétní náhodné veličiny je po částech konstantní a $F(x) = 1$ pro x větší než všechna x_i.

Každá náhodná veličina definovaná na klasickém konečném pravděpodobnostním prostoru je diskrétní.

I když hodnoty náhodné veličiny X nejsou diskrétní, můžeme postupovat podobně. Intuitivně lze při infinitesimální změně hodnoty x o přírůstek dx uvažovat takto: hustotu $f(x)$ pravděpodobnosti pro náhodnou veličinu X si představíme jako

$$P(x \leq X < x + dx) = f(x)dx.$$

To znamená, že chceme pro $-\infty \leq a \leq b \leq \infty$

$$P(a \leq X < b) = \int_a^b f(x)dx.$$

vracení) dvě koule. Jaká je pravděpodobnost, že druhá vytažená bude černá, jestliže první vytažená koule byla bílá. ○

D. Co je pravděpodobnost?

Nejprve si připomeňme geometrickou pravděpodobnost, jak jsme se s ní setkali v 1.21.

9.22. Buffonova jehla. Jehlu o délce l házíme na síť rovnoběžek tvořících pásy o šířce l. Jaká je pravděpodobnost, že jehla po dopadu zůstane v pozici protínající některou rovnoběžku?

Řešení. Pozice jehly po dopadu je dána dvěma nezávislými parametry, totiž vzdáleností d jejího středu od nejbližší rovnoběžky, ($d \in [0, l/2]$) a úhlem α ($\alpha \in [0, \pi/2]$), který jehla svírá s rovnoběžkami. Podmínka, že jehla protne některou z rovnoběžek je ekvivalentní nerovnosti $l/2 \sin\alpha > d$. Oblast možných jevů, možných dvojic (α, d), je obdélník $\pi/2 \times l/2$. Příznivé jevy, tedy ty dvojice (α, d), pro které $l/2 \sin\alpha > d$, odpovídají bodům v obdélníku ležícím pod křivkou $l/2 \sin\alpha$ (přičemž za proměnnou považujeme α, kterou vynášíme na osu x). Obsah útvaru je podle 6.35

$$\int_0^{\frac{\pi}{2}} \frac{l}{2} \sin\alpha \, \mathrm{d}\alpha = \frac{l}{2}.$$

Hledanou pravděpodobnost tak určíme (viz 1.21) jako

$$\frac{\frac{l}{2}}{\frac{\pi}{2} \cdot \frac{l}{2}} = \frac{2}{\pi}.$$

□

Následující (známý) příklad, ve kterém rovněž využijeme geometrickou pravděpodobnost, ilustruje, že si musíme dávat velký pozor na to, co považujeme za „zřejmé“

9.23. Bertrandův paradox. Určete pravděpodobnost toho, že náhodně vybraná tětiva v dané kružnici bude mít délku větší, než je strana rovnostranného trojúhelníka vepsaného do této kružnice.

Řešení. Ukážeme tři různé způsoby, jak „tuto“ pravděpodobnost odvodit.

1) Každá tětiva je jednoznačně dána svým středem. Její náhodný výběr je tedy dán náhodným výběrem jejího středu. Tětiva je delší než strana vepsaného rovnostranného trojúhelníka, leží-li její střed uvnitř soustředné kružnice o poloviční poloměru. Střed vybíráme „náhodně“ z celé kružnice, je tedy pravděpodobnost, že padne do vnitřního kruhu dána poměrem obsahů těchto kruhů, tedy je to $\frac{1}{4}$.

2) Oproti předcházejícímu odvození provedeme úvahu, že hledaná pravděpodobnost by měla být stejná, omezíme-li se pouze na tětivy

Spojité náhodné veličiny

Náhodná veličina X, pro kterou existuje její *hustota pravděpodobnosti* f splňující

$$F_X(b) = \int_{-\infty}^{b} f(x)dx,$$

se nazývá *spojitá náhodná veličina*.

Všimněme si, že distribuční funkce $F(x)$ spojité náhodné veličiny X je vždy diferencovatelná a její derivace se rovná hustotě pravděpodobnosti X, tj. platí $F'(x) = f(x)$.

Samozřejmě se také můžeme setkat se smíšeným chováním u veličin, které mají část pravděpodobnosti rozprostřenu spojitě, některých hodnot ale nabývají s nenulovou pravděpodobností. Představme si třeba chaotického přednášejícího, který s pravděpodobností p zůstává stát na místě za řečnickým pultíkem, jakmile se však odtud pohne, je jeho pozice v kterémkoliv jiném místě na stupínku stejně pravděpodobná.

Bude tedy příslušná náhodná veličina udávající jeho polohu mít distribuční funkci (zavádíme si souřadnice tak, že pultík je v pozici 0 a posluchárna je ohraničena hodnotami ± 1)

$$F(t) = \begin{cases} 0 & \text{je-li } t \le -1 \\ \frac{1-p}{2}(t+1) & \text{je-li } t \in (-1, 0) \\ p + \frac{1-p}{2}(t+1) & \text{je-li } t \in [0, 1) \\ 1 & \text{je-li } t \ge 1. \end{cases}$$

Distribuční funkce takovýchto veličin můžeme často přímo vyjadřovat pomocí Riemannova-Stieltjesova integrálu $F(t) = \int_{-\infty}^{t} f(x)d(g(x))$, který jsme zavedli v odstavci 6.48 na straně 369. V předchozím příkladu bychom zvolili třeba $f(x) = 1$ a

$$g(x) = \begin{cases} -1 & \text{pro } x \le -1 \\ \frac{1-p}{2}x & \text{pro } -1 < x < 0 \\ \frac{1-p}{2}x + p & \text{pro } 0 \le x < 1 \\ \frac{1+p}{2} & \text{pro } x \ge 1. \end{cases}$$

Připomeňme, že distribuční funkce nemůže mít více než spočetně mnoho bodů nespojitosti.

9.22. Několik diskrétních rozdělení. Požadavky na vlastnosti rozdělení náhodných veličin zpravidla vychází z modelovaných situací a ve skutečnosti pak ani nemáme moc možností, jak rozdělení pravděpodobnosti může vypadat.

Uvedeme přehled nejjednodušších diskrétních rozdělení.

Degenerované rozdělení

Rozdělení odpovídající konstantní náhodné veličině $X = \mu$ se nazývá *degenerované rozdělení* $\mathrm{Dg}(\mu)$.

Jeho distribuční funkce F_X a pravděpodobnostní funkce f_X jsou dány

$$F_X(t) = \begin{cases} 0 & t \le \mu \\ 1 & t > \mu \end{cases} \qquad f_X(t) = \begin{cases} 1 & t = \mu \\ 0 & \text{jinak} \end{cases}.$$

Nyní popišme pokus s pouze dvěma možnými výsledky, kterým budeme říkat zdar a nezdar. Pokud má zdar pravděpodobnost p, pak nezdar musí mít pravděpodobnost $1 - p$.

daného směru. Středy tětiv jednoho směru leží v dané kružnici na jediném jejím průměru daného směru. Středy vyhovujících tětiv pak jsou ty z tohoto průměru, které leží uvnitř vnitřní kružnice (viz předchozí bod), tedy na jejím průměru daného směru. Poměry průměrů kružnic jsou $1:2$, hledaná pravděpodobnost je tedy $\frac{1}{2}$.

3) Tětiva kružnice je též určena svými krajními body (ležícími na kružnici). Pokud jeden z krajních bodů tětivy, řekněme A, fixujeme (opět vzhledem k symetrii by to nemělo ovlivnit výslednou pravděpodobnost), tak druhý, aby tětiva vyhověla požadavku, musí ležet na kratším oblouku BC, kde ABC je rovnostranný trojúhelník vepsaný do dané kružnice. Délka tohoto oblouku činí jednu třetinu z obvodu kružnice. Hledaná pravděpodobnost je tedy $\frac{1}{3}$.

Jak je možné, že nám vyšla pokaždé jiná pravděpodobnost? Zadání úlohy je totiž nejednoznačné. Je nutné specifikovat, co to znamená „náhodný“ výběr tětivy. Každý ze tří pravděpodobností popisuje hledanou pravděpodobnost při vybírání tětivy popsaným způsobem. Tyto způsoby nejsou ekvivalentní, což kromě spočtené pravděpodobnosti potvrzuje i rozložení středů takto vybíraných tětiv: v prvním případě jsou středy rovnoměrně rozmístěny uvnitř celé kružnice. Ve druhém a ve třetím případě je větší koncentrace středů u středu dané kružnice.

□

9.24. Dvě obálky. V každé ze dvou obálek je umístěna určitá suma peněz. Víme, že v jedné obálce je dvojnásobek toho, co v druhé. Můžeme si zvolit jednu z obálek (a vzít si obnos v ní). Po volbě jsme dotázáni, jestli nechceme výběr změnit (a vzít si sumu z druhé obálky). Je výhodné svoje rozhodnutí změnit?

Řešení. Na první pohled musí být úplně jedno, kterou obálku zvolíme. Pravděpodobnost, že si vybereme tu s větším obnosem je $1/2$, nemá tedy smysl rozhodnutí měnit.

Proveďme však následující úvahu: v prvně zvolené obálce je suma a. Ve druhé je tedy obnos $a/2$, či $2a$, a to každý s poloviční pravděpodobností. Pokud tedy změníme své rozhodnutí, tak s pravděpodobností $1/2$ získáme obnos $a/2$, s pravděpodobností $1/2$ obnos $2a$, tedy průměrně

$$\frac{1}{2}\frac{a}{2}+\frac{1}{2}2a=\frac{5}{4}a.$$

Bylo by tedy výhodné volbu změnit. Co je špatného na této úvaze?

Je to několik věcí. Je to průměrování fiktivních výher $a/2$ a $2a$. Situace je totiž dána dvěma obálkami s výhrami a a $2a$. Při změně obálky budou naše výhry opět buď a (pokud jsme si na počátku vybrali obálku se sumou $2a$), nebo $2a$ (pokud jsme si na poprvé vybrali

Alternativní rozdělení

Rozdělení náhodné veličiny X s dvěma hodnotami 0 pro nezdar a 1 pro zdar, přičemž zdar nastává s pravděpodobností p, říkáme *alternativní rozdělení* $\mathrm{A}(p)$. Jeho distribuční a pravděpodobnostní funkce jsou tvaru:

$$F_X(t)=\begin{cases}0 & t\le 0\\ 1-p & 0<t\le 1\\ 1 & t>1\end{cases}\qquad f_X(t)=\begin{cases}p & t=1\\ 1-p & t=0\\ 0 & \text{jinak.}\end{cases}$$

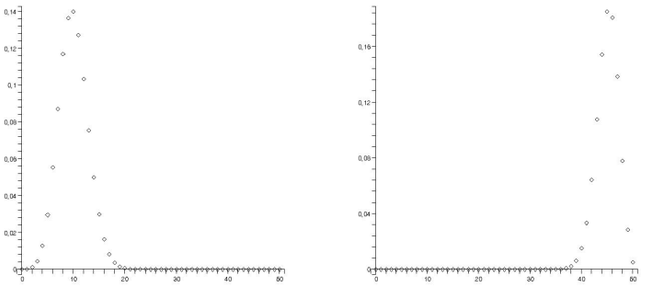
Uvažme dále rozdělení veličiny X odpovídající n–krát nezávisle opakovanému pokusu popsanému alternativním rozdělením, přičemž naše náhodná veličina měří počet zdarů. Je tedy zjevné, že pravděpodobnostní funkce bude mít nenulové hodnoty právě v celých číslech $0,\dots,n$ odpovídajícím celkovému počtu úspěchů v pokusech (a nezáleží nám na pořadí).

Binomické rozdělení

Binomické rozdělení $\mathrm{Bi}(n,p)$ má pravděpodobnostní funkci

$$f_X(t)=\begin{cases}\binom{n}{t}p^t(1-p)^{n-t} & t\in\{0,1,\dots,n\}\\ 0 & \text{jinak}\end{cases}.$$

Na obrázku jsou pravděpodobnostní funkce pro $\mathrm{Bi}(50,0{,}2)$, a $\mathrm{Bi}(50,0{,}9)$. Rozdělení pravděpodobnosti odpovídá intuici, že nejvíce výsledků bude blízko u hodnoty np:



Uvažujme nezávisle prováděné pokusy s alternativním rozdělením pravděpodobnosti $\mathrm{A}(p)$ jako u binomického rozdělení a zvolme si kladné přirozené číslo r. Budeme pokračovat v pokusech tak dlouho, dokud nenastane právě r zdarů.

Náhodná veličina X bude dána počtem nezdarů předcházejících r-tému zdaru. V případě $r=1$ tedy jsme zpět u našeho příkladu z 9.10. Náhodný jev $X=k$ nastane, právě když v prvních $k+r-1$ pokusech nastane právě $r-1$ zdarů a přitom zároveň v $(k+r)$-tém pokusu nastane zdar.

Geometrické rozdělení

Náhodná veličina X, která je dána počtem nezdarů před dosažením právě r-tého zdaru, má rozdělení pravděpodobnosti

$$P(X=k)=\binom{k+r-1}{r-1}p^r(1-p)^k,\quad k=0,1,2,\dots$$

Tomuto rozdělení že také říká *negativně binomické rozdělení*, v případě $r=1$ pak *geometrické rozdělení*.

obálku se sumou a). Tedy celková průměrná výhra je (jako na začátku)

$$\frac{1}{2}a + \frac{1}{2}2a = \frac{3}{2}a. \qquad \square$$

E. Náhodné veličina, hustota, distribuční funkce

9.25. Při jednom hodu kostkou je zřejmě množina elementárních jevů $\Omega = \{\omega_1, \ldots, \omega_6\}$, kde ω_i znamená, že na kostce padne číslo i. Jevovým polem nechť je

$$\mathcal{A} = \{\emptyset, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4, \omega_5, \omega_6\}, \Omega\}.$$

Zjistěte, jestli zobrazení $X : \Omega \to \mathbb{R}$ dané předpisem

i) $X(\omega_i) = i$ pro každé $i \in \{1, 2, 3, 4, 5, 6\}$,

ii) $X(\omega_1) = X(\omega_2) = -2, X(\omega_3) = X(\omega_4) = X(\omega_5) = X(\omega_6) = 3$

je náhodnou veličinou vzhledem k $\mathcal{A}$.

Řešení. Nejprve je dobré se přesvědčit, že množina $\mathcal{A}$ opravdu splňuje všechny axiomy v 9.11 a je tedy dobře definovaným jevovým polem. Pak podle definice 9.18 je náhodná veličina taková funkce $X : \Omega \to \mathbb{R}$, že vzor každé Borelovské množiny $B \subset \mathbb{R}$ leží v $\mathcal{A}$. Pokud v případě i) uvážíme například uzavřený interval $[2, 3]$, je jasné, že $X^{-1}([2, 3]) = \{\omega_2, \omega_3\} \notin \mathcal{A}$. Funkce X tedy v tomto případě není náhodná veličina.

V případě ii) se naopak lze lehce přesvědčit, že X je náhodná veličina. Vezmeme-li totiž libovolný interval v $\mathbb{R}$, tak mohou nastat právě čtyři možnosti. Buď neobsahuje číslo -2 ani 3, pak je vzorem X prázdná množina, pokud obsahuje jen -2, je vzorem $\{\omega_1, \omega_2\}$, pokud obsahuje jen 3, je vzorem $\{\omega_3, \omega_4, \omega_5, \omega_6\}$ a pokud interval obsahuje obě čísla, pak je vzorem celá množina Ω. Ve všech případech vzor leží v jevovém poli $\mathcal{A}$. $\square$

9.26. Je dáno jevové pole $(\Omega, \mathcal{A})$, kde $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$ a

$$\mathcal{A} = \{\emptyset, \{\omega_1, \omega_2\}, \{\omega_3\}, \{\omega_4, \omega_5\}, \{\omega_1, \omega_2, \omega_3\}, \{\omega_1, \omega_2, \omega_4, \omega_5\}, \{\omega_3, \omega_4, \omega_5\}, \Omega\}.$$

Najděte co nejobecnější zobrazení $X : \Omega \to \mathbb{R}$, které bude náhodnou veličinou vzhledem k $\mathcal{A}$.

Řešení. Protože se jevy ω_1, ω_2 nevyskytují samostatně v $\mathcal{A}$, je zřejmé, že náhodná veličina X je musí zobrazit na stejné číslo, tj. $X(\omega_1) = X(\omega_2) = a$, pro nějaké $a \in \mathbb{R}$. Ze stejného důvodu musí být $X(\omega_4) = X(\omega_5) = b$, pro nějaké $b \in \mathbb{R}$. Obsahuje-li interval obě číslo a i b, pak je jeho vzorem $\{\omega_1, \omega_2, \omega_4, \omega_5\} \in \mathcal{A}$, což je v pořádku. Zbývá jev ω_3, který se může zobrazit na libovolné $c \in \mathbb{R}$. Jednoduše se potom přesvědčíme o tom, že vzory všech intervalů

Geometrické rozdělení se ve fyzice objevuje u tzv. Einsteinovy–Boseovy statistiky.

9.23. Poissonovo rozdělení. V praktických úlohách často úvaha o binomickém rozdělení vede k dalším modelovým případům.

Uvažme situaci, kdy do n přihrádek rozdělujeme r vzájemně nerozlišitelných předmětů. Umístění kteréhokoliv předmětu do pevně zvolené přihrádky má pravděpodobnost $1/n$ (každá z nich je stejně pravděpodobná).

Náhodnou veličinu, která popisuje počet X předmětů v jedné pevně zvolené přihrádce můžeme popsat následovně. Máme možnosti hodnot $X = k$, kde $k = 0, \ldots, r$ a pravděpodobnost jednotlivých hodnot je

$$P(X = k) = \binom{r}{k}\left(\frac{1}{n}\right)^k\left(1 - \frac{1}{n}\right)^{r-k} = \binom{r}{k}\frac{(n-1)^{r-k}}{n^r}.$$

Jde proto o rozložení X typu $\mathrm{Bi}(r, 1/n)$.

S takovouto veličinou se můžeme potkat např. u popisu fyzikální soustavy s velikým počtem molekul plynu. Přihrádky představují malé objemy prostoru a sledujeme rozložení molekul. Zajímá nás pak, co se bude dít s veličinami X_n, když bude vzrůstat počet přihrádek n společně s počtem předmětů r_n tak, že v průměru nám na každou přihrádku bude připadat (přibližně) stejný počet prvků λ.

Zajímá nás tedy chování našeho rozdělení veličin X_n při limitním přechodu $n \to \infty$. Standardní úpravy (můžeme je brát i jako výzvu k opakování postupů z analýzy funkcí jedné proměnné!) vedou při $\lim_{n\to\infty} r_n/n = \lambda$ k výsledku:

$$\begin{aligned}
\lim_{n\to\infty} P(X_n = k) &= \lim_{n\to\infty}\binom{r_n}{k}\frac{(n-1)^{r_n-k}}{n^{r_n}} = \\
&= \lim_{n\to\infty}\frac{r_n(r_n-1)\ldots(r_n-k+1)}{(n-1)^k}\frac{1}{k!}\left(1-\frac{1}{n}\right)^{r_n} = \\
&= \frac{\lambda^k}{k!}\lim_{n\to\infty}\left(1+\frac{-\frac{r_n}{n}}{r_n}\right)^{r_n} = \\
&= \frac{\lambda^k}{k!}\,\mathrm{e}^{-\lambda},
\end{aligned}$$

protože obecně funkce $(1+x/n)^n$ konvergují stejnoměrně k funkci e^x na každém omezeném intervalu v $\mathbb{R}$.

Poissonovo rozdělení

Poissonovo rozdělení $\mathrm{Po}(\lambda)$ popisuje náhodné veličiny s pravděpodobnostní funkcí

$$f_X(t) = \begin{cases} \frac{\lambda^k}{k!}\,\mathrm{e}^{-\lambda} & t \in \mathbb{N} \\ 0 & \text{jinak.} \end{cases}$$

Samozřejmě platí

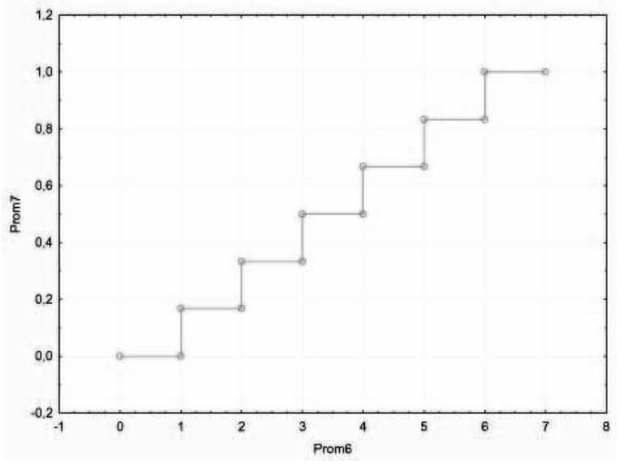
$$\sum_{k=0}^{\infty} f_X(k) = \sum_k \frac{\lambda^k}{k!}\,\mathrm{e}^{-\lambda} = \mathrm{e}^{-\lambda}\sum_k \frac{\lambda^k}{k!} = \mathrm{e}^{-\lambda+\lambda} = 1.$$

Jak jsme odvodili výše, toto diskrétní rozdělení $\mathrm{Po}(\lambda)$ s libovolným $\lambda > 0$ (rozložené do nekonečně mnoha bodů) dobře aproximuje binomická rozložení $\mathrm{Bi}(n, p_n)$, kde $np_n = \lambda$, pro veliká n.

pro takto definované X jsou množiny v $\mathcal{A}$, tj. X je náhodná veličina vzhledem k $\mathcal{A}$. □

9.27. Náhodná veličina X nabývá hodnoty i s pravděpodobností $P(X = i) = \frac{1}{6}$ pro $i = 1, \ldots, 6$. Zapište distribuční funkci $F_X(x)$ a načrtněte její graf.

Řešení. Z definice 9.19 je distribuční funkce rovna $F_X(x) = P(X < x)$. To znamená, že $F_X(x) = 0$ pro $x < 1$, $F_X(x) = \frac{[x]}{6}$ pro $1 \leq x < 6$, kde $[x]$ značí celou část čísla x, a $F_X(x) = 1$ pro $x \geq 6$. Graficky znázorněno



□

9.28. Střelec střílí do terče, dokud ho netrefí. Má v zásobě 4 náboje. Pravděpodobnost zásahu je přitom při každém výstřelu rovna 0,6. Nechť náhodná veličina X udává počet nespotřebovaných nábojů. Určete pravděpodobnostní a distribuční funkci X a nakreslete jejich grafy.

Řešení. Pravděpodobnost, že střelec k–krát terč netrefí a pak ho trefí je zřejmě rovna $0{,}4^k \cdot 0{,}6$. Proto $f_X(x) = P(X = x) = 0{,}4^{3-x} \cdot 0{,}6$ pro $x \in \{1, 2, 3\}$. Pokud střelec netrefí terč na tři pokusy, už mu každopádně nezbude žádný náboj, ať už ho v posledním pokusu trefí nebo ne. Proto $f_X(0) = P(X = 0) = 0{,}4^3$.

Z definice distribuční funkce 9.19 je

$$F_X(x) = P(X < x) = \begin{cases} 0 & \text{pro } x \leq 0, \\ 0{,}4^3 = 0{,}064 & \text{pro } x \in (0, 1], \\ 0{,}4^3 + 0{,}4^2 \cdot 0{,}6 = 0{,}16 & \text{pro } x \in (1, 2], \\ 0{,}4^3 + 0{,}4^2 \cdot 0{,}6 + 0{,}4 \cdot 0{,}6 = 0{,}4 & \text{pro } x \in (2, 3], \\ 1 & \text{pro } x > 3. \end{cases}$$

Grafy pravděpodobnostní a distribuční funkce vypadají následovně.

9.24. Dva příklady Poissonova rozdělení. Kromě výše

zmíněného fyzikálního modelu lze takové chování při sledování výskytu jevů v prostoru s konstantní očekávanou hustotou na jednotku objemu (např. při sledování výskytu bakterií na sklíčku pod mikroskopem, které se stejně pravděpodobně vyskytují v kterékoliv jeho části). Je-li „průměrná hustota výskytu“ v jednotkové ploše λ, pak při rozdělení celé oblasti na n stejných částí bude výskyt k jevů v jedné vybrané části modelován náhodnou veličinou X s Poissonovým rozdělením. Takovéto pozorování při praktické diagnostice v biochemické laboratoři umožní výpočet docela přesného celkového počtu bakterií ve vzorku ze skutečného počtu odečteného jen v několika náhodně vybraných malých částech vzorku.

Zkusme nyní popsat události, které se vyskytují náhodně v čase $t \geq 0$ a přitom pravděpodobnost výskytu v následujícím malinkém časovém intervalu o délce h nezávisí na předchozí historii a je rovna stále stejné hodnotě $h\lambda$ pro pevné $\lambda > 0$. Přitom pravděpodobnost, že nastane jev v daném malinkém intervalu více než jedenkrát bude velmi malá.

Označme si náhodnou veličinu X_t vyčíslující počet výskytu sledovaného jevu v intervalu $[0, t)$ a zkusme vyjádřit naše požadavky infinitesimálně. Chceme, aby

- pravděpodobnost právě jedné události v každém časovém úseku o délce h byla rovna $h\lambda + \alpha(h)$, kde funkce $\alpha(h)$ splňuje $\lim_{h \to 0_+} \frac{\alpha(h)}{h} = 0$;
- pravděpodobnost $\beta(h)$, že nastane více než jedna událost v časovém úseku délky h, splňuje $\lim_{h \to 0_+} \frac{\beta(h)}{h} = 0$;
- jevy $X_t = j$ a $X_{t+h} - X_t = k$ jsou nezávislé pro všechny $j, k \in \mathbb{N}$ a $t, h > 0$.

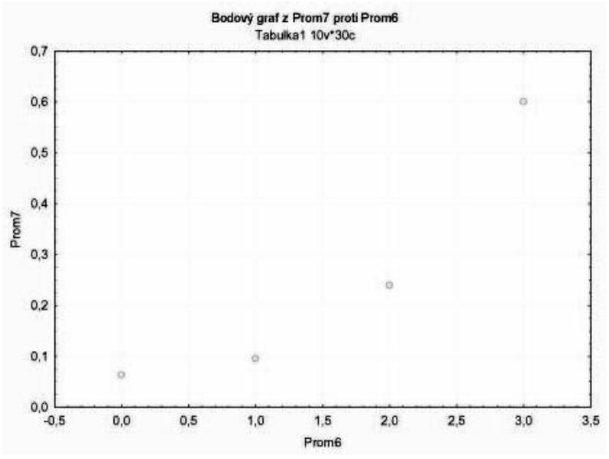
Označíme si funkce $p_k(t) = P(X_t = k)$, $k \in \mathbb{N}$, a položíme samozřejmé okrajové podmínky $p_0(0) = 1$ a $p_k(0) = 0$ pro $k > 0$. Nyní přímo spočteme

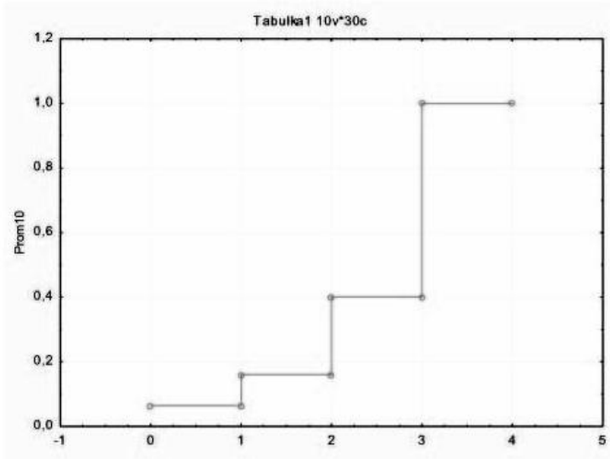
$$\begin{aligned} p_0(t + h) &= p_0(t) P(X_{t+h} - X_t = 0) = \\ &= p_0(t)(1 - h\lambda - \alpha(h) - \beta(h)) \end{aligned}$$

a podobně

$$\begin{aligned} p_k(t + h) &= P(X_t = k, X_{t+h} - X_t = 0) + \\ &\quad + P(X_t = k - 1, X_{t+h} - X_t = 1) + \\ &\quad + P(X_t \leq k - 2, X_{t+h} = k) = \\ &= p_k(t) P(X_{t+h} - X_t = 0) + p_{k-1} P(X_{t+h} - X_t = 1) + \\ &\quad + \sum_{i=0}^{k-2} P(X_t = i, X_{t+h} - X_t = k - i) = \\ &= p_k(t)(1 - h\lambda - \alpha(h) - \beta(h)) + p_{k-1}(t)(h\lambda + \alpha(h)) + \\ &\quad + \sum_{i=0}^{k-2} p_i(t) P(X_{t+h} - X_t = k - i). \end{aligned}$$

Odtud ale vidíme (píšeme stejně jako v 6.17 na straně 340 symbol $o(h)$ pro výrazy, které podělené h dávají limitu pro $h \to 0_+$





□

9.29. Náhodná veličina má distribuční funkci

$$F_X(x) = \begin{cases} 0 & \text{pro } x \le 3 \\ \frac{1}{3}x - 1 & \text{pro } 3 < x \le 6 \\ 1 & \text{pro } 6 < x. \end{cases}$$

i) Zdůvodněte, že jde skutečně o distribuční funkci.

ii) Určete hustotu pravděpodobnosti náhodné veličiny X.

iii) Vypočtěte $P(2 < X < 4)$.

Řešení. a) Jde zřejmě o spojitou neklesající funkci, která navíc vyhovuje $\lim_{x\to-\infty} F(x) = 0$ a $\lim_{x\to\infty} F(x) = 1$.

b) Podle 9.21 je hustota pravděpodobnosti spojité náhodné veličiny dána derivací distribuční funkce. Na intervalu $(3, 6)$ je tedy hustota $f(x) = \frac{1}{3}$. Na intervalech $(-\infty, 3)$ a $(6, \infty)$ je evidentně derivace nulová. Jde tedy o rovnoměrné rozdělení, viz 9.25.

c) Z definice distribuční funkce $P(2 < X < 4) = F_X(4) - F_X(2) = \frac{4}{3} - 1 = \frac{1}{3}$. □

nulovou)

$$\frac{p_0(t+h) - p_0(t)}{h} = -\lambda p_0(t) + \frac{1}{h}\,\mathrm{o}(h)$$
$$\frac{p_k(t+h) - p_k(t)}{h} = -\lambda p_k(t) + \lambda p_{k-1}(t) + \frac{1}{h}\,\mathrm{o}(h)$$

a limitním přechodem pro $h \to 0_+$ tak dostáváme (nekonečný!) systém obyčejných diferenciálních rovnic:

$$p_0'(t) = -\lambda p_0(t),\ p_0(0) = 1$$
$$p_k'(t) = -\lambda p_k(t) + \lambda p_{k-1}(t),\ p_k(0) = 0$$

pro všechny $t > 0$ a $k \in \mathbb{N}$, s počáteční podmínkou.

Nemusíme se ale děsit, protože první z nich má jediné řešení

$$p_0(t) = \mathrm{e}^{-\lambda t},$$

které okamžitě můžeme dosadit a vyřešit druhou rovnici. Obdržíme

$$p_1(t) = \lambda t\,\mathrm{e}^{-\lambda t}.$$

Matematickou indukcí teď už snadno dovodíme, že ve skutečnosti má celý systém jediné řešení a to

$$p_k(t) = \frac{(\lambda t)^k}{k!}\,\mathrm{e}^{-\lambda t},\ t > 0,\ k \in \mathbb{N}.$$

Ověřili jsme tedy, že pro každý proces splňující tři výše uvedené vlastnosti má náhodná veličina X_t udávají počet výskytů v časovém intervalu $[0, t)$ rozdělení $\mathrm{Po}(\lambda t)$.

V praxi jsou takové procesy spojeny např. s poruchovostí strojů a zařízení.

9.25. Spojitá rozdělení. Nejjednodušším příkladem spojitého

rozdělení je rovnoměrné rozprostření veškeré pravděpodobnosti na nějakém intervalu. Na něm lze dobře ilustrovat, že při jednoduše formulovaném požadavku na chování rozdělení nám nezbude moc prostoru pro jeho definici. Nyní chceme, aby pravděpodobnost hodnoty veličiny X v každém intervalu stejné délky obsaženém v daném intervalu $(a, b) \subset \mathbb{R}$ byla stejná, tj. hustota f_X našeho rozdělení náhodné veličiny X má být konstantní.

ROVNOMĚRNÉ ROZDĚLENÍ

Pro libovolná reálná čísla $-\infty < a < b < \infty$ definujeme hustotu a distribuční funkci takto:

$$f_X(t) = \begin{cases} 0 & t \le a \\ \frac{1}{b-a} & t \in (a, b) \\ 0 & t \ge b, \end{cases} \qquad F_X(t) = \begin{cases} 0 & t \le a \\ \frac{t-a}{b-a} & t \in (a, b) \\ 1 & t \ge b. \end{cases}$$

Říkáme, že veličina X má *rovnoměrné rozdělení*.

Další rozdělení budeme podobné diskrétnímu Poissonovu. Předpokládejme, že sledujeme výskyt náhodného jevu takového, že jeho výskyty v nepřekrývajících se intervalech jsou nezávislé. Je-li tedy $p(t)$ pravděpodobnost, že jev nenastane během intervalu délky t, pak nutně $p(t + s) = p(t)p(s)$ pro všechna $t, s > 0$. Předpokládejme navíc diferencovatelnost funkce p a $p(0) = 1$.

Pak jistě $\ln p(t+s) = \ln p(t) + \ln p(s)$, takže limitním přechodem (s využitím L'Hospitalova pravidla)

$$\big(\ln(p)\big)'(t) = \lim_{s\to 0_+} \frac{\ln p(t+s) - \ln p(t)}{s} = \frac{p'(0)}{p(0)} = p'(0).$$

9.30. Hustota pravděpodobnosti náhodné veličiny X má tvar $f(x) = \frac{a}{1+x^2}$ pro $x \in \mathbb{R}$. Určete

i) koeficient a,

ii) distribuční funkci,

iii) $P(-1 < X < 1)$.

Řešení. a) Aby funkce $f(x)$ byla hustotou pravděpodobnosti, musí být její integrál přes celé $\mathbb{R}$ roven jedné. Dostáváme tedy podmínku

$$1 = \int_{-\infty}^{\infty} \frac{a}{1+x^2} dx = a[\operatorname{arctg} x]_{-\infty}^{\infty} = a\pi.$$

Odtud $a = \frac{1}{\pi}$.

b) Distribuční funkce je podle 9.21 dána integrálem

$$F_X(x) = \int_{-\infty}^{x} f(t)dt = \frac{1}{\pi}\int_{-\infty}^{x} \frac{dt}{1+t^2} = \frac{1}{\pi}\operatorname{arctg} x + \frac{1}{2}.$$

c) Z definice distribuční funkce a podle b) je

$$P(-1 < X < 1) = F_X(1) - F_X(-1) = \frac{1}{\pi}\cdot\frac{\pi}{4} - \frac{1}{\pi}\cdot\left(-\frac{\pi}{4}\right) = \frac{1}{2}.$$

□

9.31. Diskrétní náhodný vektor má sdruženou pravděpodobnostní funkci danou tabulkou

X \ Y	2	5	6
1	$\frac{1}{5}$	$\frac{1}{10}$	$\frac{1}{20}$
2	$\frac{1}{10}$	$\frac{1}{20}$	0
3	$\frac{3}{10}$	$\frac{1}{20}$	$\frac{3}{20}$

Určete

i) marginální distribuční a pravděpodobnostní funkce;

ii) sdruženou distribuční funkci a vhodným způsobem ji znázorněte;

iii) $P(Y > 3X)$.

Řešení. a) Marginální rozdělení náhodné veličiny X resp. Y dostaneme podle 9.27 sečtením sdružené pravděpodobnostní funkce přes všechny možné hodnoty veličiny Y (odpovídá sečtení hodnot v každém řádku) resp. X (odpovídá sečtení hodnot v každém sloupci). Pomocí tabulky proto dostáváme

X	1	2	3
f_X	$\frac{7}{20}$	$\frac{3}{20}$	$\frac{1}{2}$

a

Y	2	5	6
f_Y	$\frac{3}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

b) Sdružená distribuční funkce v bodě (a, b) je podle definice rovna součtu všech hodnot sdružené pravděpodobnostní funkce $f_{(X,Y)}$ takových, že $X \le a$ a $Y \le b$. To v tabulce zhruba řečeno odpovídá součtu všech hodnot ležících v podtabulce, jejíž pravý spodní roh je (a, b). Přesněji máme pro sdruženou distribuční funkci $F_{(X,Y)}$ následující tabulku

Označme si proto spočtenou derivaci $p'(0) = -\lambda \in \mathbb{R}$, přičemž volíme záporné znaménko, protože víme, že $p'(0)$ nemůže být kladné, když je $p(0) = 1$.

Pak tedy pro $p(t)$ platí $\ln p(t) = -\lambda t + C$ a počáteční podmínka dává jediné řešení

$$p(t) = \mathrm{e}^{-\lambda t}.$$

Všimněme si, že z definice našich objektů vyplývá, že $\lambda > 0$.

Nyní uvažme náhodnou veličinu X udávající (náhodný) okamžik, kdy náš jev poprvé nastane. Zřejmě tedy je distribuční funkce rozdělení pro X dána

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

Je vidět, že skutečně jde o rostoucí funkci s hodnotami mezi nulou a jedničkou a správnými limitami v $\pm\infty$. Hustotu tohoto rozdělení dostaneme derivováním distribuční funkce.

EXPONENCIÁLNÍ ROZDĚLENÍ

Spojité rozdělení náhodné veličiny X s hustotou

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

se nazývá *exponenciální rozdělení* $\mathrm{ex}(\lambda)$.

Budeme také potkávat rozdělení, které je podobné jako exponenciální, ale s hustotou tvaru

$$cx^{a-1}\,\mathrm{e}^{-bx}$$

pro $x > 0$, s danými konstantami $a > 0, b > 0$, zatímco konstantu c je třeba dopočítat. Potřebujeme

$$1 = \int_0^{\infty} cx^{a-1}\,\mathrm{e}^{-bx}\,dx = \int_0^{\infty} c\left(\frac{t}{b}\right)^{a-1}\mathrm{e}^{-t}\,\frac{1}{b}\,dt = \frac{c}{b^a}\Gamma(a).$$

GAMA ROZDĚLENÍ

Rozdělení, jehož hustota je nulová pro $x \le 0$, zatímco pro $x > 0$ je dána předpisem

$$f(X) = \frac{b^a}{\Gamma(a)}x^{a-1}\,\mathrm{e}^{-bx},$$

se nazývá *gama rozdělení* $\Gamma(a, b)$ s parametry $a > 0, b > 0$.

Exponenciální rozdělení je tedy speciálním případem tohoto rozdělení s parametrem $a = 1$.

9.26. Normální rozdělení. Jestliže v binomiálním rozdělení zachováme konstantní úspěšnost p, ale budeme přidávat počet pokusů n, bude pravděpodobnostní funkce kupodivu pořád mít podobný tvar (i když jiné rozměry).

Na obrázku při rostoucím n se budou vynesené bodové hodnoty slévat do křivky, která by nám měla dát hustotu spojitého rozdělení aproximujícího dobře Bi(n, p) pro veliká n.

Naznačíme dopředu, kde hledat. Vzpomeňme na hladkou funkci $y = f(x) = \mathrm{e}^{-x^2/2}$, kterou jsme v odstavci 6.6 na straně 329 zmiňovali jako vhodný nástroj pro konstrukce funkcí hladkých, ale nikoliv analytických. Na obrázku je srovnání této křivky (vpravo) s vynesenými hodnotami Bi(5000, 0,5).

$F_{(X,Y)}$	[2,5)	[5,6)	[6,∞)
[1,2)	$\frac{1}{5}$	$\frac{3}{10}$	$\frac{7}{20}$
[2,3)	$\frac{3}{10}$	$\frac{9}{20}$	$\frac{1}{2}$
[3,∞)	$\frac{3}{5}$	$\frac{4}{5}$	1

a na intervalech $(-\infty, 1) \times \mathbb{R}$ a $\mathbb{R} \times (-\infty, 2)$ je $F_{(X,Y)}$ zřejmě nulová.
c) Očividně $P(Y > 3X) = P(X = 1, Y = 5) + P(X = 1, Y = 6) = \frac{1}{10} + \frac{1}{20} = \frac{3}{20}$ □

9.32. Určete pravděpodobnost $P(2X > Y)$, je-li hustota náhodného vektoru (X, Y)

$$f_{(X,Y)}(x, y) = \begin{cases} \frac{1}{6}(4x - y) & \text{pro } 1 \le x \le 2,\ 2 \le y \le 4, \\ 0 & \text{jinak.} \end{cases}$$

Řešení. Z definice

$$\begin{aligned} P(2X > Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{2x} f_{(X,Y)}(x, y)\,dy\,dx = \\ &= \int_1^2 \int_2^{2x} \frac{1}{6}(4x - y)\,dy\,dx = \\ &= \int_1^2 \left[\frac{2}{3}xy - \frac{1}{12}y^2\right]_2^{2x} dx = \\ &= \int_1^2 \left(x^2 - \frac{4}{3}x + \frac{1}{3}\right) dx = \\ &= \left[\frac{1}{3}x^3 - \frac{2}{3}x^2 + \frac{1}{3}x\right]_1^2 = \frac{2}{3}. \end{aligned}$$

□

9.33. Určete marginální distribuční funkce, sdruženou a marginální hustotu náhodného vektoru (X, Y), je-li

$$F_{(X,Y)}(x, y) = \begin{cases} 0 & \text{pro } x < 0,\ y < 0 \\ \frac{1}{4}x^2 y^2 & \text{pro } 0 \le x \le 1,\ 0 \le y \le 2 \\ 1 & \text{pro } x > 1,\ y > 2 \end{cases}$$

Řešení. Hustotu náhodného vektoru (X, Y) dostaneme derivováním distribuční funkce podle x a y. Tedy pro $0 \le x \le 1, 0 \le y \le 2$ je $f_{(X,Y)}(x, y) = xy$, jinde je hustota nulová. Marginální hustota náhodné veličiny X je pak

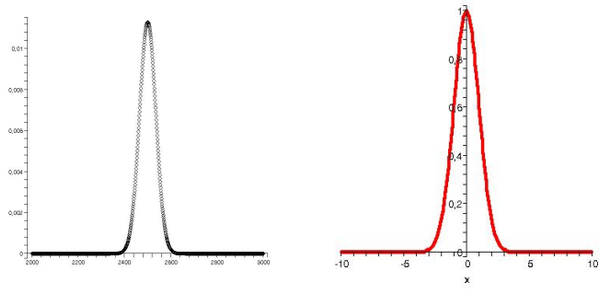
$$f_X(x) = \int_{-\infty}^{\infty} f_{(X,Y)}(x, y)\,dy = \int_0^2 xy\,dy = [\frac{1}{2}xy^2]_0^2 = 2x.$$

Podobně pro Y dostaneme $f_Y(y) = \frac{1}{2}y$. Marginální distribuční funkce jsou

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \int_0^x 2t\,dt = x^2$$

a

$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt = \int_0^y \frac{1}{2}t\,dt = \frac{1}{4}y^2.$$ □



Podbízí se proto hledat vhodné spojité rozdělení, které by mělo hustotu danou pomocí vhodně upravené takové funkce.

Funkce $e^{-x^2/2}$ je vždy kladná funkce, stačí nám spočíst $\int_{-\infty}^{\infty} e^{-x^2/2}\,dx$ a pokud to bude konečné číslo, prostě tuto funkci vynásobíme jeho převrácenou hodnotou. Spočítat tento integrál sice není možné pomocí elementárních funkcí, můžeme si ale pomoci vícerozměrnou integrací a Fubiniho větou. Snadno totiž pomocí transformace do polárních souřadnic spočteme, že

$$\begin{aligned} 2\pi &= \int_{\mathbb{R}^2} e^{-(x^2+y^2)/2}\,dx\,dy \\ &= \left(\int_{-\infty}^{\infty} e^{-x^2/2}\,dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2/2}\,dy\right) \end{aligned}$$

(srovnejte s poznámkami na konci odstavce 8.28, ověřte, že integrovaná funkce skutečně vyhovuje tam uvedeným podmínkám, a spočtěte si podrobně!). Odtud ale již vidíme, že hledaný integrál má hodnotu $\sqrt{2\pi}$ a bude proto funkce $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ dobře definovanou hustotou náhodné veličiny.

Normální rozdělení

Spojité rozdělení náhodné veličiny Z s hustotou

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

se nazývá (standardizované) *normální rozdělení* N(0, 1). Příslušnou distribuční funkci

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2}\,dx$$

nelze vyjádřit pomocí elementárních funkcí, přesto se s ní numericky běžně počítá (pomocí tabulek nebo softwarových aplikací).

Grafu hustoty $\varphi(x)$ se také často říká *Gaussova křivka*.

Tím jsme ještě evidentně nenašli tu pravou hustotu aproximující binomiální rozdělení. Obrázek, srovnávající pravděpodobnostní funkci binomického rozdělení s Gaussovou křivkou, ukazuje, že budeme chtít jednak posouvat hodnotu, kde dochází k maximální hodnotě a také zužovat či rozšiřovat oblast s výrazněji kladnými hodnotami. Toho můžeme snadno docílit vnesením dvou reálných parametrů μ a $\sigma > 0$ takto:

$$g_{\mu,\sigma}(x) = e^{-(x-\mu)^2/(2\sigma^2)}.$$

Nyní snadno spočteme pomocí jednoduché substituce proměnné

$$\int_{-\infty}^{\infty} e^{-(x-\mu)^2/(2\sigma^2)}\,dx = \sqrt{2\pi}\sigma.$$

Dostáváme tedy celou dvouparametrickou třídu hustot

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

9.34. V urně je 14 kuliček – 4 červené, 5 bílých a 5 modrých. Náhodně bez vracení vybereme 6 kuliček. Určete rozložení náhodného vektoru (X, Y), označuje-li X počet tažených červených kuliček a Y počet tažených bílých kuliček. Určete rovněž marginální rozložení veličin X a Y. Dále vypočtěte $P(X \leq 3)$, $P(1 \leq Y \leq 4)$.

Řešení. Z definice pravděpodobnostní funkce je její hodnota v bodě (x, y) určena pravděpodobností $P(X = x, Y = y)$, tedy pravděpodobností, že vytáhneme x červených a y bílých kuliček. Počet možností vytažení x Červených kuliček je $\binom{4}{x}$, podobně počet vytažení y bílých je $\binom{5}{y}$. Zbylých $6 - x - y$ modrých kuliček pak můžeme vytáhnout $\binom{5}{6-x-y}$ způsoby. Dohromady tedy máme $\binom{4}{x}\binom{5}{y}\binom{5}{6-x-y}$ možností. Hodnoty tohoto výrazu pro všechny možné hodnoty x, y jsou v následující tabulce.

x\ y	0	1	2	3	4	5	$\sum_X$
0	0	5	50	100	50	5	210
1	4	100	400	400	100	4	1008
2	30	300	600	300	30	0	1260
3	40	200	200	40	0	0	480
4	10	25	10	0	0	0	45
$\sum_Y$	84	630	1260	840	180	9	3003

Hodnoty nejvíce napravo a dole jsou součty možností přes všechny hodnoty y resp. x. Hodnoty pravděpodobnostní funkce jsou pak zřejmě dány vydělením příslušných hodnot v tabulce počtem všech výběrů šesti kuliček, tj. vydělením číslem $\binom{14}{6} = 3003$. Marginální rozložení X resp. Y je přitom dáno hodnotami nejvíce napravo resp. dole v tabulce.

Pravděpodobnost $P(X \leq 3)$ jednoduše spočítáme pomocí marginálního rozložení X

$$P(X \leq 3) = F_X(3) = \frac{1}{3003}(210 + 1008 + 1260 + 480) = 0{,}985.$$

Podobně pro pravděpodobnost $P(1 \leq Y \leq 4)$ máme

$$P(1 \leq Y \leq 4) = F_Y(4) - F_Y(1) = \\ = \frac{1}{3003}(630 + 1260 + 840 + 180) = 0{,}969.$$

□

9.35. Hustota náhodného vektoru (X, Y, Z) je

$$f(x, y, z) = \begin{cases} c(x + y + z) & \text{pro } 0 \leq x \leq 1, 0 \leq y \leq 1, 0 \leq z \leq 1 \\ 0 & \text{jinak.} \end{cases}$$

Určete konstantu c, distribuční funkci a vypočtěte

$$P(0 \leq X \leq \tfrac{1}{2}, 0 \leq Y \leq \tfrac{1}{2}, 0 \leq Z \leq \tfrac{1}{2}).$$

náhodných veličin. Příslušná rozdělení budeme značit $\mathrm{N}(\mu, \sigma)$.

K asymptotické blízkosti normálního a binomického rozdělení pro $n \to \infty$ se ještě vrátíme, jenom co si k tomu vytvoříme patřičné nástroje.

9.27. Rozdělení náhodných vektorů. Obdobně jako u skalárních veličin definujeme distribuční funkce a hustotu nebo pravděpodobnostní funkci pro spojité a diskrétní náhodné vektory. Hovoříme také o simultánních (sdružených) pravděpodobnostních funkcích a hustotách.

Pro dvě diskrétní náhodné veličiny, tj. diskrétní vektor (X, Y) náhodných veličin, definujeme *(sdruženou) pravděpodobnostní funkci*

$$f(x, y) = \begin{cases} P(X = x_i \wedge Y = y_j) & x = x_i,\ y = y_j \\ 0 & \text{jinak.} \end{cases}$$

Pro spojité veličiny pak definujeme pro všechny $a, b \in \mathbb{R}$

$$F(a, b) = P(X < a, Y < b) = \\ = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x, y) dx dy$$

a funkci $f(x, y)$ nazýváme *(sdruženou) hustotou* náhodného vektoru (X, Y).

Pro obecný náhodný vektor $X = (X_1, \ldots, X_n)$ obdobně při spojitých náhodných veličinách X_i definujeme

$$F(a_1, \ldots, a_n) = P(X_1 < a_1, \ldots, X_n < a_n) = \\ = \int_{-\infty}^{a_n} \cdots \int_{-\infty}^{a_1} f(x, y)\, dx_1 \cdots dx_n$$

a obdobně pro diskrétní náhodné veličiny.

Marginální rozložení pro jednu z proměnných obdržíme tak, že přes ostatní posčítáme nebo zintegrujeme.

Např. u diskrétních vektorových veličin (X, Y) tvoří jevy $(X = x_i, Y = y_j)$ pro všechny možné hodnoty x_i a y_j s nenulovými pravděpodobnostmi pro X a Y úplný systém jevů pro vektor (X, Y) a dostáváme vztah:

$$P(X = x_i) = \sum_{j=1}^{\infty} P(X = x_i, Y = y_j)$$

mezi *marginálním rozdělením pravděpodobnosti* náhodné veličiny X a sdruženým rozdělením pravděpodobnosti náhodného vektoru (X, Y). Zcela obdobně postupujeme u spojitých náhodných vektorů s pomocí integrálů.

9.28. Nezávislost náhodných veličin. Víme už, co to je (ne)závislost náhodných jevů, kterou jsme diskutovali v odstavci 9.12. O náhodných veličinách $X_1, \cdots, X_n$ řekneme, že jsou *(stochasticky) nezávislé*, jestliže jsou pro libovolná čísla $a_i \in \mathbb{R}$ nezávislé jevy $X_1 < a_1, \ldots, X_n < a_n$.

To již díky axiomům pravděpodobnosti zaručuje, že budou nezávislé i všechny jevy zadané příslušností hodnot veličin $X_k \in I_k$ do libovolných intervalů I_k. Přímo z definičních vlastností pak také odvodíme, že jsou náhodné veličiny X_i nezávislé, právě když jejich sdružená distribuční funkce F splňuje

$$F(x_1, \ldots, x_n) = F_1(x_1) \cdots F_n(x_n),$$

kde F_i jsou distribuční funkce veličin X_i.

Řešení. Integrál hustoty pravděpodobnosti přes celý prostor musí být roven jedné, a proto

$$1 = \int_0^1 \int_0^1 \int_0^1 c(x+y+z)dzdydx = c\int_0^1 \int_0^1 (x+y+\tfrac{1}{2})dydx = \\ = c\int_0^1 (x+1)dx = \tfrac{3}{2}c.$$

Odtud $c = \frac{2}{3}$. Distribuční funkce je z definice rovna

$$F_X(x,y,z) = \tfrac{2}{3}\int_0^x \int_0^y \int_0^z (r+s+t)dtdsdr = \\ = \tfrac{2}{3}\int_0^x \int_0^y (rz+sz+\tfrac{1}{2}z^2)dsdr = \tfrac{2}{3}\int_0^x (rzy+\tfrac{1}{2}y^2z+\tfrac{1}{2}z^2y)dr = \\ = \tfrac{2}{3}(\tfrac{1}{2}x^2zy+\tfrac{1}{2}y^2zx+\tfrac{1}{2}z^2yx) = \tfrac{1}{3}(x^2zy+y^2zx+z^2yx),$$

a proto je hledaná pravděpodobnost

$$P(0 \le X \le \tfrac{1}{2}, 0 \le Y \le \tfrac{1}{2}, 0 \le Z \le \tfrac{1}{2}) = F(\tfrac{1}{2},\tfrac{1}{2},\tfrac{1}{2}) = \tfrac{1}{16}. \qquad \square$$

9.36. Určete konstantu a tak aby funkce

$$f(x) = \begin{cases} 0 & \text{pro } x \le 1 \\ a\ln(x) & \text{pro } 1 < x < 2 \\ 0 & \text{pro } 2 \le x \end{cases}$$

zadávala hustotu pravděpodobnosti nějaké náhodné veličiny.

Řešení. Podmínka na to, aby zadaná funkce zadávala hustotu pravděpodobnosti je

$$\int_{-\infty}^{\infty} f(x) = 1$$

Bude potřeba spočítat $\int \ln(x)\,\mathrm{d}x$:

$$\int \ln(x)\,\mathrm{d}x = x\ln(x) - \int 1\,\mathrm{d}x = x\ln(x) - x = x(\ln(x)-1).$$

Celkem

$$\int_{-\infty}^{\infty} f(x) = \int_1^2 a\ln(x) = a[x(\ln(x)-1)]_1^2 = a(2\ln(2)-1),$$

tedy $a = \frac{1}{2\ln(2)-1}$. $\square$

9.37. V lese, jehož hranice tvoří na mapě pravidelný šestiúhelník se ztratilo dítě. Předpokládejme, že pravděpodobnost toho, že dítě je v určité části lesa, je úměrná pouze velikosti této části, nikoliv jejímu umístění.

- Jaké je rozdělení pravděpodobnosti vzdálenosti dítěte od zvolené strany (přímky) lesa
- Jaké je rozdělení pravděpodobnosti vzdálenosti dítěte od nejbližší strany lesa.

Řešení.

- Nechť a je strana šestiúhelníka. Pak rozdělení pravděpodobnosti je

$$f(x) = \begin{cases} 0 & \text{pro } x \le 0 \\ \frac{4}{9a^2}x + \frac{2}{3\sqrt{3}a} & \text{pro } 0 < x \le \frac{1}{2}\sqrt{3}a \\ -\frac{4}{9a^2}x + \frac{2}{\sqrt{3}a} & \text{pro } \frac{1}{2}\sqrt{3}a \le x \le \sqrt{3}a \\ 0 & \text{pro } x > \sqrt{3}a \end{cases},$$

pro první část.

Pro diskrétní nezávislé veličiny z definice okamžitě vyplývá, že sdružená pravděpodobnostní funkce nezávislých veličin je dána součiny jednotlivých hodnot

$$f_{X,Y}(x,y) = \sum_{x_i}\sum_{y_j} x_i y_j.$$

Derivací sdružené distribuční funkce spojitých proměnných dostáváme obdobný vztah mezi jejich hustotami:

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) = \\ = \frac{\partial^2}{\partial x \partial y} F_X(x) F_Y(y) = \\ = f_X(x) f_Y(y).$$

Jde tedy o prostý součin hustot jednotlivých veličin.

Hustoty náhodných vektorů vyšších dimenzí s nezávislými spojitými komponentami se chovají zcela obdobně a jejich sdružené hustoty jsou součinem hustot jednotlivých veličin, tj.

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = f_{X_1}(x_1)\cdots f_{X_n}(x_n).$$

Podívejme se na jednoduchý příklad, který ukazuje, že není dobré zjednodušeně vidět náhodný vektor, coby stochastický objekt, jen jako dvojici náhodných veličin. Uvažme náhodný vektor (X,Y), který má rovnoměrné spojité rozdělení na jednotkovém kruhu v rovině $\mathbb{R}^2$ se středem v počátku. Bude tedy jeho (sdružená) hustota

$$f(x,y) = \begin{cases} \frac{1}{\pi} & x^2+y^2 \le 1 \\ 0 & \text{jinak.} \end{cases}$$

Je vidět, že veličiny X a Y z tohoto náhodného vektoru nemohou být nezávislé. Např. si všimněme, že pravděpodobnost, že (X,Y) padne do doplňku jednotkového kruhu ve čtverci s vrcholy o souřadnicích ± 1 je nulová, zatímco marginální distribuční funkce jsou pro hodnoty $|x| \le 1$ a $|y| \le 1$ nenulové.

Když ovšem vyjádříme tentýž náhodný vektor v polárních souřadnicích (R,Φ), dostáváme

$$P(R < r_0, \Phi < \varphi_0) = \int_0^{r_0}\int_0^{\varphi_0} \frac{1}{\pi} r\,d\varphi dr = \frac{1}{2}\varphi_0 r_0^2.$$

Sdružená hustota vektoru (R,Φ) je tedy $f(r,\varphi) = \frac{r}{\pi}$ při $0 < r \le 1, 0 < \varphi \le 2\pi$ a jinak je nulová. Marginální hustoty jsou

$$f_R(r) = \int_0^{2\pi} \frac{r}{\pi}\,d\varphi = 2r, \quad \text{je-li } 0 < r \le 1,$$

$$f_\Phi(\varphi) = \int_0^1 \frac{r}{\pi}\,dr = \frac{1}{2\pi}, \quad \text{je-li } 0 < \varphi \le 2\pi,$$

a nula jinak. Náhodné veličiny R a Φ jsou tedy nezávislé.

9.29. Funkce náhodných veličin. 
Náhodné vektory potkáváme v praktických modelech ve dvou velmi odlišných rolích. Můžeme sledovat skutečně několik různých náhodných veličin popisujících více či méně související jevy. Jako příklad nám mohou sloužit různorodé číselné parametry svázané s jednotlivými studenty (prospěch v různých předmětech, váha, výška, stáří, roční příjem, atd.). V tomto případě budeme potřebovat nástroje, které nám umožní sledovat rozdíly či závislosti mezi takovými veličinami.

Můžeme ale také sledovat jen jeden parametr na velkém souboru objektů a vybíráme přitom jen menší počet n z nich. Takový

- Spočtěme nejprve distribuční funkci F hledaného rozložení náhodné veličiny X udávající vzdálenost dítěte od nejbližší strany lesa. Vzdálenost se může pohybovat v intervalu $I = \langle 0, \frac{\sqrt{3}}{2}a\rangle$. Pro $y \in I$ potom máme

$$F(y) = P[X < y] = \frac{\frac{\sqrt{3}}{4}a^2 - \frac{(\frac{\sqrt{3}}{2}a-y)^2}{\frac{3}{4}a^2}\frac{\sqrt{3}}{4}a^2}{\frac{\sqrt{3}}{4}a^2} = 1 - \frac{4(\frac{\sqrt{3}}{2}a-y)^2}{3a^2}$$

Celkem tedy

$$F(y) = \begin{cases} 0 & \text{pro } y \le 0 \\ 1 - \frac{4(\frac{\sqrt{3}}{2}a-y)^2}{3a^2} & \text{pro } y \in \langle 0, \frac{\sqrt{3}}{2}a\rangle \\ 1 & \text{pro } y \ge \frac{\sqrt{3}}{2}a \end{cases}$$

Pro hustotu pravděpodobnosti, která je derivací distribuční funkce dostáváme:

$$f(x) = \begin{cases} 0 & \text{pro } x \le 0 \\ \frac{8(\frac{\sqrt{3}}{2}a-y)}{3a^2} & \text{pro } y \in \langle 0, \frac{\sqrt{3}}{2}a\rangle \\ 0 & \text{pro } y \ge \frac{\sqrt{3}}{2}a \end{cases}$$

□

9.38. Nechť veličina náhodná veličina X má rovnoměrné rozdělení na intervalu $\langle 0, r\rangle$. Určete distribuční funkci a hustotu pravděpodobnosti rozdělení objemu koule o poloměru X.

Řešení. Určeme nejprve distribuční funkci F (pro $0 < d < \frac{4}{3}\pi r^3$)

$$F(d) = P\left[\frac{4}{3}\pi X^3 \le d\right] = P\left[X \le \sqrt[3]{\frac{3d}{4\pi}}\right] = \frac{\sqrt[3]{\frac{3d}{4\pi}}}{r},$$

celkem

$$F(x) = \begin{cases} 0 & \text{pro} \quad x \le 0 \\ \sqrt[3]{\frac{3}{4\pi r^3}}\, x^{\frac{1}{3}} & \text{pro} \quad 0 < x < \frac{4}{3}\pi r^3 \\ 1 & \text{pro} \quad x \ge \frac{4}{3}\pi r^3 \end{cases}$$

Derivováním pak obdržíme hustotu pravděpodobnosti:

$$f(x) = \begin{cases} 0 & \text{pro} \quad x \le 0 \\ \sqrt[3]{\frac{1}{36\pi r^3}}\, x^{-\frac{2}{3}} & \text{pro} \quad 0 < x < \frac{4}{3}\pi r^3 \\ 0 & \text{pro} \quad x \ge \frac{4}{3}\pi r^3 \end{cases}$$

□

9.39. Stanovte hodnotu parametru $a \in \mathbb{R}$ tak, aby funkce

$$f(x) = \begin{cases} 0 & \text{pro} \quad x \le 0 \\ ax^2 & \text{pro} \quad 0 < x < 3 \\ 0 & \text{pro} \quad x \ge 3 \end{cases}$$

zadávala hustotu pravděpodobnosti náhodné veličiny X. Určete distribuční funkci, hustotu pravděpodobnosti a střední hodnotu rozdělení objemu krychle, jejíž délka hrany je náhodná veličina s hustotou pravděpodobnosti danou funkcí f.

postup popisujeme pomocí n–rozměrného vektoru $(X_1, \ldots, X_n)$, kde všechny náhodné veličiny X_k mají stejné rozdělení pravděpodobnosti. Tady nás budou velice zajímat veličiny, které budou odpovídat statistickým číselným charakteristikám, které jsme již potkali v předchozí části této kapitoly.

Budeme umět oba případy zvládat jedním jednoduchým konceptem. Místo dané náhodné veličiny nebo náhodného vektoru budeme uvažovat funkci z těchto veličin.

I u jediné veličiny jde o velice užitečný nástroj. Místo náhodné veličiny X, např. „roční plat zaměstnance“, budeme vyčíslovat jinou závislou hodnotu $\psi(X)$, např. „roční čistý příjem zaměstnance po zdanění a včetně sociálních dávek“. V systému s tzv. sociální solidaritou je první veličina hodně variabilní, zatímco druhá může být skoro konstantní. Statisticky se proto budou značně odlišovat.

Funkce náhodných veličin a vektorů

Pro danou spojitou funkci $\psi : \mathbb{R} \to \mathbb{R}$ a náhodnou veličinu X máme dánu také náhodnou veličinu $Y = \psi(X)$. Nazýváme ji *funkcí náhodné veličiny* X.

V případě náhodného vektoru $(X_1, \ldots, X_n)$ a funkce $\psi : \mathbb{R}^n \to \mathbb{R}$ hovoříme o funkci $Y = \psi(X_1, \ldots, X_n)$ náhodného vektoru.

Všimněme si, že požadavek spojitosti ψ zaručuje, že je Y opět náhodnou veličinou podle naší definice, protože vzor borelovské množiny ve spojitém zobrazení je opět borelovská množina. Obecněji můžeme právě tento požadavek na ψ vztáhnout pro každý speciální případ veličiny či vektoru a definovat tak pojem funkce z náhodné veličiny či vektoru obecněji.

Nejjednodušší funkcí po konstantách je afinní závislost

$$\psi(X) = a + bX$$

s konstantami $a, b \in \mathbb{R}$, $b \neq 0$.

Je-li $f_X(x)$ pravděpodobnostní funkce náhodné veličiny s diskrétním rozdělením, snadno se vypočte

$$f_{\psi(X)}(y) = P(\psi(X) = y) = \sum_{\psi(x_i)=y} f(x_i).$$

V případě afinní závislosti $Y = a+bX$ je proto pravděpodobnostní funkce nenulová právě v bodech $y_i = ax_i + b$.

Jako příklad na funkci náhodného vektoru si rozmyslete součet n nezávislých náhodných veličin s alternativním rozdělením $\mathrm{A}(p)$. Samozřejmě dostáváme právě binomiální rozdělení $\mathrm{Bi}(n, p)$.

Podobně můžeme přepočíst distribuční funkci rozdělení funkce ze spojité náhodné veličiny, či vektoru. Ukážeme na příkladu.

V předposledním odstavci jsme zavedli veličinu Z s normálním rozdělením $\mathrm{N}(0, 1)$. Snadno spočteme, že veličiny $Y = \mu + \sigma Z$ budou mít normální rozdělení $\mathrm{N}(\mu, \sigma)$ diskutované tamtéž. Skutečně,

$$\begin{aligned} F_Y(y) &= P(Y < y) = P(\mu + \sigma Z < y) = \\ &= \Phi\big(\frac{1}{\sigma}(y-\mu)\big) = \tfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{y-\mu}{\sigma}} \mathrm{e}^{-z^2/2}\, dz = \\ &= \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}\sigma}\, \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx, \end{aligned}$$

kde jsme v posledním kroku použili substituci $x = \mu + \sigma z$. To je právě požadovaný výraz.

Řešení. Jednoduše $a = \frac{1}{9}$. Distribuční funkce náhodné veličiny X je tedy $F_X(t) = \frac{1}{27}t^3$ pro $t \in (0, 3)$, pro menší t je tato funkce nulová, pro větší rovna 1. Označme $Z = X^3$ náhodnou veličinu označující objem krychle. Ten je v intervalu $(0, 27)$, pro $t \in (0, 27)$ a distribuční funkci F_Z náhodné veličiny Z tedy můžeme psát $F_Z(t) = P[Z < t] = P[X^3 < t] = P[X < \sqrt[3]{t}] = F_X(\sqrt[3]{t}) = \frac{1}{27}t$, hustota pravděpodobnosti je pak $f_Z(t) = \frac{1}{27}$ na intervalu $(0, 27)$, jinak nula, jedná se tedy o rovnoměrné rozdělení pravděpodobnosti na daném intervalu, střední hodnota je tudíž 13, 5. □

9.40. Stanovte hodnotu parametru $a \in \mathbb{R}$ tak, aby funkce

$$f(x) = \begin{cases} 0 & \text{pro } x \le 0 \\ ax & \text{pro } 0 < x < 3 \\ 0 & \text{pro } x \ge 3 \end{cases}$$

zadávala hustotu pravděpodobnosti náhodné veličiny X. Určete distribuční funkci, hustotu pravděpodobnosti a střední hodnotu rozdělení obsahu čtverce, jehož délka hrany je náhodná veličina s hustotou pravděpodobnosti danou funkcí f.

Řešení. Budeme postupovat jako v předchozím příkladě. Opět snadno zjistíme $a = \frac{2}{9}$. Distribuční funkce náhodné veličiny X je tedy $F_X(t) = \frac{1}{9}t^2$ pro $t \in (0, 3)$, pro menší t je tato funkce nulová, pro větší rovna 1. Označme $Z = X^2$ náhodnou veličinu označující obsah čtverce. Ten je v intervalu $(0, 9)$, pro $t \in (0, 9)$ a distribuční funkci F_Z náhodné veličiny Z tedy můžeme psát $F_Z(t) = P[Z < t] = P[X^2 < t] = P[X < \sqrt{t}] = F_X(\sqrt{t}) = \frac{1}{9}t$, hustota pravděpodobnosti je pak $f_Z(t) = \frac{1}{9}$ na intervalu $(0, 9)$, jinak nula, jedná se tedy o rovnoměrné rozdělení pravděpodobnosti na daném intervalu, střední hodnota je tudíž 4, 5. □

9.41. Stanovte hodnotu parametru $a \in \mathbb{R}$ tak, aby funkce

$$f(x) = \begin{cases} 0 & \text{pro } x \le 0 \\ ax^2 & \text{pro } 0 < x < 2 \\ 0 & \text{pro } x \ge 2 \end{cases}$$

zadávala hustotu pravděpodobnosti náhodné veličiny X. Určete distribuční funkci, hustotu pravděpodobnosti a střední hodnotu rozdělení objemu krychle, jejíž délka hrany je náhodná veličina s hustotou pravděpodobnosti danou funkcí f. ○

9.42. Náhodně rozřízneme úsečku délky l na dvě části. Určete distribuční funkci a hustotu pravděpodobnosti rozdělení obsahu obdélníka, jehož délky stran jsou rovny délkám takto vzniklých úseček.

Řešení. Spočítejme hledanou distr. funkci. Označme ještě X náhodnou veličinu s rovnoměrným rozložením na intervalu $\langle 0, l \rangle$ udávající délku jedné ze stran (délka druhé je pak $l - X$). Obsah obdélníka

Se součty nezávislých náhodných veličin je to malinko složitější. Uvažme dvě takové spojité veličiny X a Y s hustotami f_X a f_Y. Přímým výpočtem spočteme distribuční funkci náhodné proměnné $V = X + Y$.

$$F_V(u) = \int_{x+y<u} f_X(x) f_Y(y)\, dxdy = \\ = \int_{-\infty}^{u} \left(\int_{-\infty}^{\infty} f_X(x) f_Y(v - x)\, dx \right) dv.$$

Je tedy sdruženou hustotou součtu dvou nezávislých veličin právě konvoluce jejich hustot

$$f_V = f_X * f_Y,$$

se kterou jsme se setkali již v odstavci 7.28 na straně 425. Úplně stejně dostaneme diskrétní konvoluci pravděpodobnostních funkcí v případě diskrétních náhodných veličin.

Konvoluci jsme si v 7. kapitole představovali jako jisté „rozmlžení" hodnot jedné z funkcí pomocí jádra vyjádřeného druhou. Promyslete si, že to je ta správná intuice i pro hustotu součtu nezávislých náhodných veličin. Je proto také samozřejmé, že má být konvoluce symetrická vůči oběma argumentům.

9.30. Číselné charakteristiky náhodných veličin. Viděli jsme, že při statistickém zkoumání hodnot (např. zpracování výsledků nějakého měření) hledáme výpovědi pomocí číselných charakteristik, jako jsou aritmetický průměr a směrodatná odchylka. Nyní zavedeme obdobné charakteristiky pro náhodné veličiny a náhodné vektory. První z nich je obdobou aritmetického průměru.

STŘEDNÍ HODNOTA

Pro libovolnou náhodnou veličinu X definujeme její *střední hodnotu* $\mathrm{E}\,X$ vztahem

$$\mathrm{E}\,X = \begin{cases} \sum_i x_i f_X(x_i) & \text{pro diskrétní veličinu} \\ \int_{-\infty}^{\infty} x f_X(x) dx & \text{pro spojitou veličinu,} \end{cases}$$

pokud uvedené sumy či integrály absolutně konvergují. Když absolutně nekonvergují, říkáme, že náhodná veličina X střední hodnotu nemá.

Střední hodnotou náhodného vektoru rozumíme vektor středních hodnot jeho jednotlivých komponent.

Střední hodnotu můžeme přímo vyjádřit také pro funkce $Y = \psi(X)$ náhodné veličiny či vektoru X. Připomeňme, že uvažujeme pouze takové funkce ψ, kdy je Y opět náhodnou veličinou.

V diskrétním případě můžeme přímo spočíst

$$\mathrm{E}\,Y = \sum_j y_j P(Y = y_j) = \\ = \sum_j y_j \sum_{\psi(x_i)=y_j} P(X = x_j) = \\ = \sum_i \psi(x_i) P(X = x_i),$$

pokud suma absolutně konverguje. Samozřejmě, není zaručeno, že funkce z náhodné veličiny, které má střední hodnotu, bude mít střední hodnotu také.

S, tedy součin $x(l-x)$ pro $x \in \langle 0, l\rangle$ může zřejmě nabývat hodnot $\langle 0, l^2/4\rangle$. Volíme-li $d \in \langle 0, l^2/4\rangle$, můžeme psát

$$F(d) = P[S \le d] = P[X(l-X) \le d]$$

Hledáme tedy ty hodnoty x, pro které je $x(l-x) \le d$. Řešíme kvadr. nerovnici, kořeny odpovídající kvadratické rovnice jsou $\frac{l-\sqrt{l^2-4d}}{2}$ a $\frac{l+\sqrt{l^2-4d}}{2}$, hodnoty x uvnitř tohoto intervalu nerovnici nesplňují, hodnoty vně potom ano. Je tedy

$$P[X(l-X) \le d] = P[X \in \langle 0, l\rangle \setminus (\frac{l-\sqrt{l^2-4d}}{2}, \frac{l+\sqrt{l^2-4d}}{2})]$$
$$= \frac{l-\sqrt{l^2-4d}}{l} = 1 - \frac{\sqrt{l^2-4d}}{l}$$

Celkem

$$F(x) = \begin{cases} 0 & \text{pro} \quad x \le 0 \\ 1 - \frac{\sqrt{l^2-4x}}{l} & \text{pro} \quad 0 \le x \le \frac{l^2}{4} \\ 1 & \text{pro} \quad x > \frac{l^2}{4} \end{cases}$$

Hustotu pravděpodobnosti pak dostaneme derivací:

$$f(x) = \begin{cases} 0 & \text{pro} \quad x \le 0 \\ \frac{2}{l\sqrt{l^2-4x}} & \text{pro} \quad 0 \le x \le \frac{l^2}{4} \\ 0 & \text{pro} \quad x > \frac{l^2}{4} \end{cases}$$

□

9.43. Nezávislé náhodné veličiny X a Y mají následující hustoty pravděpodobnosti:

$$f_X(t) = \begin{cases} 0 & \text{pro } t \le 0, \\ 1 & \text{pro } 0 < t < 1, \\ 0 & \text{pro } 1 \le t, \end{cases} \quad f_Y(t) = \begin{cases} 0 & \text{pro } t \le 0, \\ 2t & \text{pro } 0 < t < 1, \\ 0 & \text{pro } 1 \le t. \end{cases}$$

Určete distribuční funkci náhodné veličiny udávající obsah obdélníka o stranách X a Y.

Řešení.

$$F_Y(t) = \begin{cases} 0 & \text{pro} \quad t \le 0 \\ 2t - t^2 & \text{pro} \quad 0 < t < 1 \\ 1 & \text{pro} \quad 1 \le t \end{cases}$$

□

9.44. Nechť X, Y jsou nezávislé náhodné veličiny, přičemž X má rovnoměrné rozdělení pravděpodobnosti na intervalu $(0, 2)$, Y je pak dána následující hustotou pravděpodobnosti:

$$f(x) = \begin{cases} 0 & \text{pro} \quad x \le 0 \\ 2x & \text{pro} \quad 0 < x < 1 \\ 0 & \text{pro} \quad x \ge 1. \end{cases}$$

Určete pravděpodobnost, že Y je menší než X^2.

Podobně vyjádříme střední hodnotu funkce ze spojité náhodné veličiny:

$$\mathrm{E}\,\psi(X) = \int_{-\infty}^{\infty} \psi(x) f_X(x) dx,$$

pokud tento integrál absolutně konverguje.

Všimněme si náhodná veličina $Y = \psi(X)$ nemusí být pro funkci spojité náhodné veličiny opět spojitá. Nicméně v případě spojité monotónní funkce ψ a spojité veličiny X tomu tak bude a mělo by být vcelku jednoduchým cvičením ověřit, že námi definovaná $\mathrm{E}\,\psi(X)$ skutečně splývá s $\mathrm{E}\,Y$.

Střední hodnotu náhodné veličiny můžeme nahlížet jako „očekávanou hodnotu“. Ve statistice zanedlouho uvidíme, že má skutečně přímý vztah k aritmetickému průměru vektoru hodnot.

9.31. Petrohradský paradox. Vraťme se k příkladu, kterým jsme motivovali potřebu diskrétních náhodných veličin v odstavci 9.10. Přeformulujeme tentýž model jako potenciální pravidla herny a dostaneme pěkný příklad situace, ve které střední hodnota zkoumané veličiny nebude podle naší definice vůbec existovat.

Návštěvník zaplatí vklad C a poté hází mincí. Je-li T počet hodů potřebných k první hlavě, pak obdrží výhru 2^T. Ptáme se, jaká je „rozumná hodnota“ pro vklad C? Je-li X náhodná veličina popisující výhru, jistě se nám zdá, že správnou odpovědí je „cokoliv menší než střední hodnota $\mathrm{E}\,X$“.

Jak jsme odvodili v 9.10, je (za předpokladu férové mince) $P(T = k) = 2^{-k}$. Sečteme-li všechny pravděpodobnosti výsledků vynásobené výhrami 2^k, dostaneme $\sum_1^\infty 1 = \infty$. Střední hodnota tedy neexistuje. Zdá se proto, že se hráči vyplatí vložit i velký vklad...

Ve skutečnosti simulací hry zjistíme, že nezávisle na počtu pokusů se prakticky všechny výhry budou pohybovat v rozmezí cca do 2^4. Důvodem je, že nikdo nemůže hrát neomezeně dlouho a vysoké výhry jsou proto velice nepravděpodobné a proto je při reálných úvahách nelze brát vážně. V teorii rozhodování se takovým případům, kdy očekávaná hodnota nemá přímý vztah k vyčíslenému užitku říká *Petrohradský paradox* a k této tématice lze najít rozsáhlou literaturu.[3]

9.32. Vlastnosti střední hodnoty. U jednoduchých rozdělení můžeme snadno spočíst jejich střední hodnotu přímo z definice. Např. pro náhodnou veličinu s alternativním rozdělením $\mathrm{A}(p)$ spočteme okamžitě

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

Stejně tak bychom mohli spočíst střední hodnotu np binomického rozdělení $\mathrm{Bi}(n, p)$, to už ale dá trochu přemýšlení. Nicméně výsledek je okamžitým důsledkem následující obecné věty, protože $\mathrm{Bi}(n, p)$ je součtem n náhodných veličin s alternativním rozděleními $\mathrm{A}(p)$.

Uvažme nějaké náhodné veličiny X, Y, reálné konstanty a, b a podívejme se na střední hodnoty funkcí veličin $X + Y$ a $a + bX$, za předpokladu, že střední hodnoty $\mathrm{E}\,X$ a $\mathrm{E}\,Y$ existují.

Přímo z definice je samozřejmé, že konstantní náhodná veličina a má za střední hodnotu opět a. Dále,

$$\mathrm{E}(bX) = b\,\mathrm{E}\,X,$$

protože konstanta b se vytkne jak ze sum, tak z integrálů.

[3] Bernoulli, 1738, viz Wiki – hodnota není dána cenou ale užitkem

Řešení. Protože X a Y jsou nezávislé náhodné veličiny, je sdružená hustota pravděpodobnosti $f_{(X,Y)} : \mathbb{R}^2 \to \mathbb{R}^2$ veličiny (X, Y) dána součinem hustot pravděpodobnosti f_X veličiny X a f_Y veličiny Y, tedy

$$f_{(X,Y)}(u, v) = \\ = \begin{cases} f_X(u) \cdot f_Y(v) = \frac{1}{2} \cdot 2v = v & \text{pro } (u, v) \in (0, 2) \times (0, 1), \\ 0 & \text{jinak.} \end{cases}$$

Hledaná pravděpodobnost P je pak dána integrálem hustoty pravděpodobnosti $f_{(X,Y)}$ přes tu část roviny O, kde je $Y < X^2$:

$$P = \iint_O f_{(X,Y)} \,\mathrm{d}x\,\mathrm{d}y = 1 - \iint_{\mathbb{R}^2 \setminus O} f_{(X,Y)} \,\mathrm{d}x\,\mathrm{d}y = \\ = 1 - \int_0^1 \int_{x^2}^1 y \,\mathrm{d}y\,\mathrm{d}x = \frac{3}{5}.$$

□

9.45. Nechť X, Y jsou nezávislé náhodné veličiny, přičemž X je dána následující hustotou pravděpodobnosti:

$$f_1(x) = \begin{cases} 0 & \text{pro } x \le 0 \\ 2x & \text{pro } 0 < x < 1 \\ 0 & \text{pro } x \ge 1, \end{cases}$$

veličina Y pak touto hustotou pravděpodobnosti:

$$f_2(x) = \begin{cases} 0 & \text{pro } x \le 0 \\ \frac{x}{2} & \text{pro } 0 < x < 2 \\ 0 & \text{pro } x \ge 2. \end{cases}$$

Určete pravděpodobnost, že Y je větší než X^2. ○

Řešení. $f_{(X,Y)}(u, v) = uv$, pro $(u, v) \in (0, 1) \times (0, 2)$, $f_{(X,Y)}(u, v) = 0$ jinak. Pro hledanou pravděpodobnost P pak máme

$$P = \int_0^1 \int_{x^2}^2 xy \,\mathrm{d}y\,\mathrm{d}x = \frac{11}{12}.$$ □

9.46. Nechť X, Y jsou nezávislé náhodné veličiny, přičemž X je dána následující hustotou pravděpodobnosti:

$$f(x) = \begin{cases} 0 & \text{pro } x \le 0 \\ \frac{2x}{9} & \text{pro } 0 < x < 3 \\ 0 & \text{pro } x \ge 1, \end{cases}$$

veličina Y pak touto hustotou pravděpodobnosti:

$$f(x) = \begin{cases} 0 & \text{pro } x \le 0 \\ \frac{x}{2} & \text{pro } 0 < x < 2 \\ 0 & \text{pro } x \ge 2. \end{cases}$$

Určete pravděpodobnost, že Y je větší než X^3.

○

Řešení.

$$P = \int_0^{\sqrt[3]{2}} \int_{x^3}^2 xy \,\mathrm{d}y\,\mathrm{d}x = \frac{\sqrt[3]{4}}{12}.$$ □

Obecněji, střední hodnotu součinu dvou nezávislých náhodných veličin X a Y spočteme následovně. Předpokládejme, že vektor (X, Y) má diskrétní nezávislé komponenty s pravděpodobnostními funkcemi $f_X(x_i)$, $f_Y(y_j)$. Potom

$$\mathrm{E}(XY) = \sum_i \sum_j x_i y_j f_X(x_i) f_Y(y_j) = \\ = \Big(\sum_i x_i f_X(x_i)\Big)\Big(\sum_j y_j f_Y(y_j)\Big) = \mathrm{E}\,X\ \mathrm{E}\,Y.$$

Podobně se spočte rovnost $\mathrm{E}(XY) = \mathrm{E}\,X\ \mathrm{E}\,Y$ pro nezávislé spojité veličiny.

Zkusme nyní spočíst $\mathrm{E}(X+Y)$ pro jakékoliv náhodné veličiny. Pro diskrétní rozdělení X a Y dostaneme

$$\mathrm{E}(X + Y) = \sum_i \sum_j (x_i + y_j) P(X = x_i, Y = y_j) = \\ = \sum_i \Big(x_i \sum_j P(X = x_i, Y = y_j) \Big) + \\ + \sum_j \Big(y_j \sum_i P(X = x_i, Y = y_j) \Big) = \\ = \sum_i x_i P(X = x_i) + \sum_j y_j P(Y = y_j),$$

přičemž absolutní konvergence první dvojité sumy vyplývá z trojúhelníkové nerovnosti a absolutní konvergence sum pro střední hodnotu jednotlivých proměnných, při výpočtu jsme pak absolutní konvergence sum využili k záměně pořadí sčítání.

Podobně budeme postupovat u spojitých náhodných veličin X a Y se střední hodnotou. Připomeňme, že hustota součtu náhodných veličin je dána konvolucí jejich hustot.

$$\mathrm{E}(X + Y) = \int_{-\infty}^{\infty} z(f_X * f_Y)(z) dz = \\ = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z f_X(x) f_Y(z - x)\, dx dz = \\ = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (z - x) f_X(x) f_Y(z - x)\, dx dz + \\ + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_X(x) f_Y(z - x)\, dx dz = \\ = \int_{-\infty}^{\infty} u f_Y(u)\, du + \int_{-\infty}^{\infty} x f_X(x)\, dz,$$

kde jsme využili absolutní konvergenci integrálů středních hodnot $\mathrm{E}\,X$ a $\mathrm{E}\,Y$ k záměně integrálů dle Fubiniho věty.

Celkem tedy dostáváme očekávaný vztah:

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

kdykoliv střední hodnoty $\mathrm{E}\,X$ a $\mathrm{E}\,Y$ existují.

Nyní již přímým použitím tohoto vztahu dostáváme:

AFINNÍ POVAHA STŘEDNÍ HODNOTY

Pro jakékoliv konstanty $a, b_1, \ldots, b_k$ a náhodné veličiny $X_1, \ldots, X_k$ platí

$$\mathrm{E}(a + b_1 X_1 + \cdots + b_k X_k) = a + b_1\,\mathrm{E}\,X_1 + \cdots + b_k\,\mathrm{E}\,X_k.$$

F. Střední hodnota, korelace

Spočtěte střední hodnotu a rozptyl binomického rozdělení

Řešení. Přímý výpočet z definic je pěkné kombinatorické cvičení. My tvrzení dokážeme s využitím vlastnosti středních hodnot a rozptylu. Podle definice binomického rozdělení v 9.22 můžeme náhodnou veličinu $X \sim \mathrm{Bi}(n,p)$ vidět jako součet $X = \sum_{k=1}^{n} Y_k$, kde $Y_1, \ldots, Y_n \sim A(p)$ jsou nezávislé náhodné veličiny vyjadřující úspěch v k–tém pokusu. Alternativní rozdělení má zřejmě střední hodnotu $\mathrm{E}\,Y_i = p$, a proto podle věty 9.32 platí $\mathrm{E}\,X = \sum_{k=1}^{n} \mathrm{E}\,Y_k = np$. Podobně snadno vypočteme $\mathrm{E}(Y_k^2) = 1^2 \cdot p + 0^2 \cdot (1-p) = p$, a proto $\mathrm{var}\,Y_k = \mathrm{E}(Y_k^2) - (\mathrm{E}\,Y_k)^2 = p - p^2$. Podle věty 9.36 pak platí $\mathrm{var}\,X = \sum_{k=1}^{n} \mathrm{var}\,Y_k = np(1-p)$. □

9.47. Pravděpodobnost zásahu cíle jedním výstřelem je $0,6$. Náhodná veličina X udává počet zásahů při pěti nezávislých výstřelech. Určete její rozdělení pravděpodobnosti, střední hodnotu a rozptyl.

Řešení. Výstřely jsou zřejmě nezávislé pokusy s alternativním rozdělením $A(\frac{3}{5})$, a proto je podle definice binomického rozdělení $X \sim \mathrm{Bi}(5, \frac{3}{5})$. Podle ‖F‖ je střední hodnota a rozptyl $\mathrm{Bi}(n,p)$ rovna np respektive $np(1-p)$, což v našem případě dává $\mathrm{E}\,X = 3$ a $\mathrm{var}\,X = \frac{6}{5}$. □

9.48. Diskrétní náhodná veličina X nabývá hodnot $k = 0, 1, 2, 3, \ldots$ s pravděpodobností $P(X = k) = p(1-p)^k$ (geometrické rozdělení). Určete $\mathrm{E}\,X$ (střední doba čekání na úspěch) a $\mathrm{var}\,X$.

Řešení. Z definice střední hodnoty a s využitím formule pro součet derivace geometrické řady spočítáme

$$\mathrm{E}\,X = \sum_{k=0}^{\infty} kp(1-p)^k = p(1-p)\sum_{k=0}^{\infty} k(1-p)^{k-1} = \\ = p(1-p)\frac{1}{p^2} = \frac{1-p}{p}.$$

Obdobně s využitím formule pro součet druhé derivace geometrické řady spočítáme

$$E(X^2) \quad = \quad \sum_{k=0}^{\infty} k^2 p(1-p)^k = \frac{(1-p)(2-p)}{p^2}$$

a proto je rozptyl roven $\mathrm{var}\,X = \mathrm{E}(X^2) - (\mathrm{E}\,X)^2 = \frac{1-p}{p^2}$. □

9.49. Náhodná veličina X má hustotu $f_X(x) = \frac{3}{x^4}$ pro $x \in (1, \infty)$ a jinde nulovou. Určete její distribuční funkci, střední hodnotu a rozptyl.

Následující věta rozšiřuje toto chování vůči afinním transformacím na náhodné vektory a ukazuje, že je střední hodnota invariantní vůči afinním transformacím, stejně jako aritmetický průměr:

Věta. *Nechť $X = (X_1, \ldots, X_n)$ je náhodný vektor se střední hodnotou $\mathrm{E}\,X$, $a \in \mathbb{R}^m$, $B \in \mathrm{Mat}_{mn}(\mathbb{R})$ matice. Pak platí*

$$\mathrm{E}(a + B \cdot X) = a + B \cdot \mathrm{E}\,X.$$

Důkaz. Ve skutečnosti už skoro nemáme co dokazovat. Protože je střední hodnota vektoru definována jako vektor středních hodnot, stačí se nám omezit na jedinou položku v $\mathrm{E}(a + B \cdot X)$. Můžeme proto rovnou předpokládat, že a je skalár a B matice s jediným řádkem. Pak jde ovšem o střední hodnotu konečného součtu náhodných veličin a ta podle předchozí úvahy jednak existuje a zároveň je dána jako součet středních hodnot jednotlivých položek. To je právě dokazovaný vztah. □

9.33. Kvantily a kritické hodnoty. Pokračujeme v našem programu zavádění číselných charakteristik v obdobě k těm z popisné statistiky. Dalšími užitečnými charakteristikami tam byly tzv. *kvantily*.

Uvažme nejprve náhodnou veličinu s ryze monotonní distribuční funkcí F_X. Podmínce vyhovuje každá spojitá náhodná veličina X se všude nenulovou hustotou, jako je tomu např. u normálního rozdělení. V tomto případě definujeme tzv. *kvantilovou funkci* F_X^{-1} prostě jako inverzní funkci $(F_X)^{-1} : (0,1) \to \mathbb{R}$. To znamená, že hodnota $y = F^{-1}(\alpha)$ je právě takové y, že $P(X < y) = \alpha$. To přesně odpovídá kvantilům z popisné statistiky, když budeme za pravděpodobnosti brát relativní četnosti výskytu hodnot.

Kvantilová funkce

Obecně, pro libovolnou náhodnou veličinu X s distribuční funkcí $F_X(x)$ definujeme její *kvantilovou funkci*

$$F^{-1}(\alpha) = \inf\{x \in \mathbb{R};\ F(x) \geq \alpha\},\ \alpha \in (0,1).$$

Zřejmě jde o zobecnění předchozí definice v případě ryze monotónní distribuční funkce.

Jak jsme viděli v popisné statistice, nejčastěji jsou používané kvantily s $\alpha = 0{,}5$, tzv. *medián*, s $\alpha = 0{,}25$, tzv. *první kvartil*, $\alpha = 0{,}75$, tzv. *třetí kvartil*, a podobně pro *decily* a *percentily* (kdy je α rovno násobkům desetin a setin).

Jak vyplývá přímo z definice, kvantilová funkce nám pro danou náhodnou veličinu X umožňuje přímo určovat intervaly, do kterých nám padnou hodnoty X s předem zadanou pravděpodobností. Velice často se budeme potkávat např. s hodnotou $\Phi^{-1}(0{,}975)$, která je přibližně rovna 1,96 a zadává percentil 97,5 pro normální rozdělení $\mathrm{N}(0,1)$. Tato hodnota říká, že s 2,5–procentní pravděpodobností bude hodnota takové náhodné veličiny Z alespoň 1,96. Protože je přitom hustota pravděpodobnosti veličiny Z symetrická kolem počátku, můžeme toto pozorování interpretovat tak, že pouze s 5–procentní pravděpodobností bude hodnota $|Z|$ větší než 1,96.

Podobné intervaly a hodnoty budeme hledat při diskusi spolehlivosti odhadů hodnot charakteristik náhodných veličin.

Řešení. Z definice distribuční funkce je pro $x \in (1, \infty)$

$$F_X(x) = \int_1^x \frac{3}{t^4} dt = \left[-\frac{1}{t^3}\right]_1^x = 1 - \frac{1}{x^3}.$$

Střední hodnota X je rovna

$$\mathrm{E}\, X = \int_1^\infty \frac{3}{x^3} dx = \left[-\frac{3}{2x^2}\right]_1^\infty = \frac{3}{2}$$

a střední hodnota X^2 je

$$\mathrm{E}(X^2) = \int_1^\infty \frac{3}{x^2} dx = \left[-\frac{3}{x}\right]_1^\infty = 3.$$

Proto $\operatorname{var} X = 3 - (\frac{3}{2})^2 = \frac{3}{4}$. □

9.50. Náhodná veličina X má hustotu rovnu $f_X(x) = \cos x$ pro $x \in \langle 0, \frac{\pi}{2}\rangle$ a jinde nulovou. Určete střední hodnotu, rozptyl a medián této veličiny.

Řešení. Z definice a integrací per partes spočítáme

$$\mathrm{E}\, X = \int_0^{\frac{\pi}{2}} x \cos x dx = [x \sin x + \cos x]_0^{\frac{\pi}{2}} = \frac{\pi}{2} - 1.$$

Dvojitou integrací per partes dostaneme

$$\mathrm{E}(X^2) = \int_0^{\frac{\pi}{2}} x^2 \cos x dx =$$
$$= \left[x^2 \sin x + 2x \cos x - 2 \sin x\right]_0^{\frac{\pi}{2}} = \left(\frac{\pi}{2}\right)^2 - 2,$$

a proto je rozptyl roven $\operatorname{var} X = (\frac{\pi}{2})^2 - 2 - (\frac{\pi}{2} - 1)^2 = \pi - 3$. Distribuční funkce je podle definice rovna $F_X(x) = \int_0^x \cos t dt = \sin x$ a medián $F^{-1}(0{,}5) = \frac{\pi}{6}$. □

9.51. Náhodná veličina X má hustotu rovnu $f_X(x) = \lambda e^{-\lambda x}$ pro $x \geq 0$, kde $\lambda > 0$ je daný parametr rozdělení, a jinde nulovou (tzv. exponenciální rozdělení). Určete střední hodnotu, rozptyl, modus (reálné číslo s maximální hustotou, resp. pravděpodobnostní funkcí) a medián této veličiny.

Řešení. Z definice a integrací per partes

$$\mathrm{E}\, X = \int_0^\infty x\lambda e^{-\lambda x} dx = \left[-xe^{-\lambda x} - \frac{1}{\lambda} e^{-\lambda x}\right]_0^\infty = \frac{1}{\lambda},$$
$$\mathrm{E}(X^2) = \int_0^\infty x^2 \lambda e^{-\lambda x} dx =$$
$$= \left[-x^2 e^{-\lambda x} - 2x \frac{1}{\lambda} e^{-\lambda x} - \frac{2}{\lambda^2} e^{-\lambda x}\right]_0^\infty = \frac{2}{\lambda^2},$$

a proto $\operatorname{var} X = \mathrm{E}(X^2) - (\mathrm{E}\, X)^2 = \frac{1}{\lambda^2}$. Protože $F'_X(x) = -\lambda^2 e^{-\lambda x} < 0$, je hustota stále klesající funkce. Své maximum tedy nabývá v nule. Z definice je

$$F(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}$$

a proto je medián roven $F^{-1}(0{,}5) = -\frac{1}{\lambda} \ln(\frac{1}{2}) = \frac{\ln 2}{\lambda}$. □

KRITICKÉ HODNOTY

Pro náhodnou veličinu X a reálné číslo $0 < \alpha < 1$ definujeme její *kritickou hodnotu $x(\alpha)$ na úrovni α* předpisem

$$P(X \geq x(\alpha)) = \alpha.$$

To znamená, že $x(\alpha) = F_X^{-1}(1-\alpha)$, kde F_X^{-1} je kvantilová funkce veličiny X.

9.34. Rozptyl a směrodatná odchylka. Nejjednodušší číselné charakteristiky udávající variabilitu hodnot vzorku v popisné statistice byly rozptyl a směrodatná odchylka. Pro náhodné veličiny si budeme počínat obdobně.

ROZPTYL NÁHODNÉ VELIČINY

Pro náhodnou veličinu X s konečnou střední hodnotou definujeme její *rozptyl* vztahem

$$\operatorname{var} X = \mathrm{E}\left((X - \mathrm{E}\, X)^2\right),$$

pokud i střední hodnota na pravé straně výrazu existuje. V opačném případě říkáme, že veličina X nemá rozptyl.

Odmocnina $\sqrt{\operatorname{var} X}$ z rozptylu se nazývá *směrodatná odchylka náhodné veličiny X*.

S využitím vlastností střední hodnoty snadno spočteme jednodušší vztah pro rozptyl náhodné veličiny X se střední hodnotou:

$$\operatorname{var} X = \mathrm{E}(X - \mathrm{E}\, X)^2 = \mathrm{E}(X^2 - 2X(\mathrm{E}\, X) + (\mathrm{E}\, X)^2) =$$
$$= \mathrm{E}\, X^2 - 2(\mathrm{E}\, X)^2 + (\mathrm{E}\, X)^2 =$$
$$= \mathrm{E}\, X^2 - (\mathrm{E}\, X)^2.$$

Podívejme se také, jak se chová rozptyl náhodné veličiny při afinních transformacích. Pro náhodnou veličinu X se střední hodnotou a rozptylem a pro reálná čísla a, b uvažujme náhodnou veličinu $Y = a + bX$. Spočteme

$$\operatorname{var} Y = \mathrm{E}\big((a + bX) - \mathrm{E}(a + bX)\big)^2 == \mathrm{E}\big(b(X - \mathrm{E}\, X)\big)^2$$
$$= b^2 \operatorname{var} X.$$

Odvodili jsem tedy následující užitečné vztahy:

VLASTNOSTI ROZPTYLU

$$\operatorname{var} X = E(X^2) - (EX)^2 \tag{9.8}$$
$$\operatorname{var}(a + bX) = b^2 \operatorname{var} X \tag{9.9}$$
$$\sqrt{\operatorname{var}(a + bX)} = b\sqrt{\operatorname{var} X} \tag{9.10}$$

Ke každé náhodné veličině X se střední hodnotou a rozptylem můžeme zadat tzv. *normovanou veličinu* (často také říkáme *standardizovanou veličinu*) jako funkci

$$Z = \frac{X - \mathrm{E}\, X}{\sqrt{\operatorname{var} X}}.$$

Je to tedy taková afinní transformace původní veličiny, která má střední hodnotu nulovou a rozptyl jednotkový.

9.52. Diskrétní náhodný vektor (X_1, X_2) má simultánní pravděpodobnostní funkci $\pi(0,-1) = c, \pi(1,0) = \pi(1,1) = \pi(2,1) = 2c, \pi(2,0) = 3c$ a rovnou nule jinde. Určete konstantu c a vypočtěte kovarianci $\operatorname{cov}(X_1, X_2)$.

Řešení. Součet pravděpodobnostních funkcí přes všechny možné stavy musí být roven 1, tj.

$$\sum_{i,j} \pi(i,j) = c + 3.2c + 3c = 10c = 1,$$

a odtud $c = \frac{1}{10}$. Pravděpodobnostní funkce π_1 pro X_1 je dána součtem simultánní funkce přes všechny možné hodnoty X_2 , tj. $\pi_1(i) = \sum_j \pi(i,j)$. Je tedy rovna $\pi_1(0) = c$, $\pi_1(1) = 4c$, $\pi_1(2) = 5c$ a nule jinde. Podobně pro pravděpodobnostní funkci π_2 náhodné veličiny X_2 dostaneme $\pi_2(-1) = c$, $\pi_2(0) = 5c$, $\pi_2(1) = 4c$ a nula jinde. Odtud $\mathrm{E}\,X_1 = \sum_i i\pi_1(i) = 14c = 1{,}4$ a $\mathrm{E}\,X_2 = \sum_j j\pi_2(j) = 3c = 0{,}3$. Z definice kovariance pak máme

$$\operatorname{cov}(X_1, X_2) = \sum_{i,j} (i - 1{,}4)(j - 0{,}3)\pi(i,j) = 0{,}18. \qquad \square$$

9.53. V mnoha vědních oborech se chování náhodné proměnné omezené na nějaký interval modeluje pomocí tzv. beta rozdělení. Toto spojité rozdělení je dáno pravděpodobnostní funkcí na intervalu $[0, 1]$

$$f_X(x) = \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1},$$

kde α, β jsou vhodně zvolené parametry pro popis dané náhodné veličiny a $\mathrm{B}(\alpha,\beta)$ je normalizační konstanta, která zajišťuje, že integrál $f_X(x)$ přes celý interval $[0, 1]$ je roven jedné. Spočítejte jeho a) modus, b) střední hodnotu a c) rozptyl.

Řešení. a) Modus je z definice hodnota, ve které nabývá funkce $f_X(x)$ své maximum. Hledejme tedy její stacionární body. Jednoduše spočítáme, že rovnice $f'_X(x) = 0$ je ekvivalentní rovnici

$$(\alpha - 1)(1 - x) - x(\beta - 1) = 0,$$

která je splněna pro $x = \frac{\alpha-1}{\alpha+\beta-2}$. Protože $f_X(0) = f_X(1) = 0$ a funkce je kladná, jedná se evidentně o hledané maximum.
b) Z definice je

$$\mathrm{E}\,X = \frac{1}{\mathrm{B}(\alpha,\beta)} \int_0^1 x^{\alpha}(1-x)^{\beta-1}\,dx.$$

Integrací per partes pak dostáváme

$$\mathrm{E}\,X = -\frac{1}{\mathrm{B}(\alpha,\beta)\beta}[x^{\alpha}(1-x)^{\beta}]_0^1 + \frac{\alpha}{\mathrm{B}(\alpha,\beta)\beta} \int_0^1 x^{\alpha-1}(1-x)^{\beta}\,dx.$$

9.35. Čebyševova nerovnost. Hezkou ilustrací, k čemu je užitečný rozptyl, je skoro samozřejmá nerovnost, která dává přímo do souvislosti pravděpodobnost vzdálenosti hodnot náhodné veličiny od její střední hodnoty.

ČEBYŠEVOVA NEROVNOST

Věta. *Předpokládejme, že náhodná veličina X má konečný rozptyl, a uvažujme libovolné $\varepsilon > 0$. Potom platí*

$$P(|X - \mathrm{E}\,X| \geq \varepsilon) \leq \frac{\operatorname{var} X}{\varepsilon^2}.$$

DŮKAZ. Uvedeme jednoduchý důkaz pro spojitou náhodnou veličinu X. Analogický postup pro diskrétní veličiny ponecháme na čtenáři.

Označme si $\mu = \mathrm{E}\,X$ a počítejme podle definice

$$\begin{aligned}\operatorname{var} X &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx = \\ &= \int_{|x-\mu|\geq\varepsilon} (x-\mu)^2 f(x)\,dx + \\ &\quad + \int_{|x-\mu|<\varepsilon} (x-\mu)^2 f(x)\,dx \geq \\ &\geq \int_{|x-\mu|\geq\varepsilon} \varepsilon^2 f(x)\,dx = \varepsilon^2\, P(|X-\mu| \geq \varepsilon). \end{aligned} \qquad \square$$

Když si uvědomíme, že rozptyl je kvadrát směrodatné odchylky σ, tak okamžitě vidíme, že volba $\varepsilon = k\sigma$ dává pravděpodobnost

$$P(|X - \mathrm{E}\,X| \geq k\sigma) \leq \frac{1}{k^2}.$$

Čebyševova nerovnost je mimořádně užitečná pro asymptotické odhady u limitních procesů. Uvažme např. posloupnost náhodných veličin $X_1, X_2, \ldots$ s rozložením pravděpodobnosti $X_n \sim \mathrm{Bi}(n, p)$ se stejným $0 < p < 1$. Asi bychom intuitivně očekávali, že relativní četnost zdaru by se měla s rostoucím n blížit pravděpodobnosti p, tj. že náhodné veličiny $Y_n = \frac{1}{n}X_n$ by se měl stále více svými hodnotami blížit p. Evidentně máme

$$\mathrm{E}\,Y_n = \frac{np}{n} = p, \quad \operatorname{var} Y_n = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

Přímé použití Čebyševovy nerovnosti dává pro libovolné pevné $\varepsilon > 0$

$$P\big(|Y_n - p| \geq \varepsilon\big) \leq \frac{p(1-p)}{n\varepsilon^2}.$$

Odtud ale je zřejmé, že pro každé pevně $\varepsilon > 0$ platí

$$\lim_{n\to\infty} P\Big(\Big|\frac{X_n}{n} - p\Big| \geq \varepsilon\Big) = 0.$$

Tento výsledek je známý jako *Bernoulliova věta* (jedna z mnoha).

Tomuto typu limitního chování říkáme *konvergence podle pravděpodobnosti*. Dokázali jsme tedy, že v důsledku Čebyševovy nerovnosti konvergují naše veličiny Y_n podle pravděpodobnosti ke konstantní veličině p.

9.36. Kovariance. Vraťme se nyní k náhodným vektorům. U střední hodnoty jsme to měli snadné – uvažovali jsme prostě vektor středních hodnot. Pro charakterizaci variability nás však také moc zajímají závislosti mezi jednotlivými komponentami.

První člen je očividně nulový. Úpravou druhého pak dostáváme

$$\mathrm{E}\,X = \frac{\alpha}{\mathrm{B}(\alpha,\beta)\beta}\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx - \\ -\frac{\alpha}{\mathrm{B}(\alpha,\beta)\beta}\int_0^1 x^{\alpha}(1-x)^{\beta-1}\,dx.$$

Nyní integrál v prvním členu je díky normalizaci roven právě $\mathrm{B}(\alpha,\beta)$ a druhý integrál udává též střední hodnotu. Předchozí rovnici tedy můžeme zapsat ve tvaru

$$\mathrm{E}\,X = \frac{\alpha}{\beta} - \frac{\alpha}{\beta}\,\mathrm{E}\,X.$$

Odtud okamžitě $\mathrm{E}\,X = \frac{\alpha}{\alpha+\beta}$.

c) Pro výpočet rozptylu potřebujeme spočítat

$$\mathrm{E}\,X^2 = \frac{1}{\mathrm{B}(\alpha,\beta)}\int_0^1 x^{\alpha+1}(1-x)^{\beta-1}\,dx.$$

Tento integrál spočítáme podobným způsobem jako v b). Konkretně

$$\mathrm{E}\,X^2 = \frac{\alpha+1}{\mathrm{B}(\alpha,\beta)\beta}\int_0^1 x^{\alpha}(1-x)^{\beta}dx = \frac{\alpha+1}{\beta}\,\mathrm{E}\,X - \frac{\alpha+1}{\beta}\,\mathrm{E}\,X^2.$$

Odtud $\mathrm{E}\,X^2 = \frac{(\alpha+1)\,\mathrm{E}\,X}{\alpha+\beta+1}$. Dosazením střední hodnoty pak máme

$$\operatorname{var} X = \mathrm{E}\,X^2 - (\mathrm{E}\,X)^2 = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}. \qquad \square$$

9.54. Hodíme třemi mincemi. Určete korelační koeficient veličiny X udávající počet padlých líců dohromady na první a druhé minci a veličiny Y udávající počet padlých líců dohromady na druhé a třetí minci.

Řešení. Nejprve sestavíme pravděpodobnostní tabulku vektorové diskrétní náhodné veličiny (X, Y), ze které snadno určíme pravděpodobnostní rozdělení veličin, které budeme potřebovat (samozřejmě to můžeme udělat i bez tabulky):

X \ Y	0	1	2
0	$\frac{1}{8}$	$\frac{1}{8}$	0
1	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$
2	0	$\frac{1}{8}$	$\frac{1}{8}$

Diskrétní veličiny X a Y mají stejné rozdělení pravděpodobnosti a to hodnotu 0 nabývají s pravděpodobností 1/4, hodnotu 1 s pravděpodobností 1/2 a hodnotu 2 s pravděpodobností 1/4. Veličina XY pak může nabývat hodnot 0, 1, 2, 4 a to postupně s pravděpodobnostmi 3/8, 1/4, 1/4, 1/8 Nyní spočítáme střední hodnoty veličin X, X^2, Y,

Kovariance

Pro náhodné veličiny X, Y s existujícími rozptyly definujeme jejich *kovarianci* předpisem

$$\operatorname{cov}(X,Y) = \mathrm{E}\,((X-\mathrm{E}\,X)(Y-\mathrm{E}\,Y))$$

Velmi snadno odvodíme základní vlastnosti tohoto pojmu:

Věta. *Pro jakékoliv náhodné veličin X, Y, Z, pro které existují jejich rozptyly, a reálná čísla a, b, c, d platí*

$$\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X) \tag{9.11}$$

$$\operatorname{cov}(X,Y) = \mathrm{E}(XY) - (\mathrm{E}\,X)(\mathrm{E}\,Y) \tag{9.12}$$

$$\operatorname{cov}(X+Y,Z) = \operatorname{cov}(X,Z) + \operatorname{cov}(Y,Z) \tag{9.13}$$

$$\operatorname{cov}(a+bX, c+dY) = bd\operatorname{cov}(X,Y) \tag{9.14}$$

$$\operatorname{var}(X+Y) = \operatorname{var} X + \operatorname{var} Y + 2\operatorname{cov}(X,Y). \tag{9.15}$$

Jsou-li navíc naše veličiny X a Y nezávislé, pak $\operatorname{cov}(X,Y) = 0$. *Zejména potom platí*

$$\operatorname{var}(X+Y) = \operatorname{var} X + \operatorname{var} Y. \tag{9.16}$$

Důkaz. Symetrie kovariance v argumentech je okamžitě vidět z definice . Druhé tvrzení ihned plyne z vlastností střední hodnoty náhodné veličiny:

$$\begin{aligned}\operatorname{cov}(X,Y) &= \mathrm{E}(X-\mathrm{E}\,X)(Y-\mathrm{E}\,Y) = \\ &= \mathrm{E}(XY) - (\mathrm{E}\,Y)X - (\mathrm{E}\,X)Y + \mathrm{E}\,X\,\mathrm{E}\,Y = \\ &= \mathrm{E}(XY) - \mathrm{E}\,X\,\mathrm{E}\,Y\end{aligned}$$

I další tvrzení vyplývá z rozepsání definičního vztahu a skutečnosti, že střední hodnota součtu náhodných veličin je součet jejich středních hodnot.

Další tvrzení opět také spočteme přímo:

$$\begin{aligned}&\operatorname{cov}(a+bX, c+dY) = \\ &\quad= \mathrm{E}\big((a+bX-\mathrm{E}(a+bX))(c+dY-\mathrm{E}(c+dY))\big) = \\ &\quad= \mathrm{E}\big((bX-b\,\mathrm{E}(X))(dY-d\,\mathrm{E}(Y))\big) = \\ &\quad= \mathrm{E}\big(bd(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\big) = \\ &\quad= bdE\big((X-\mathrm{E}\,X)(Y-\mathrm{E}\,Y)\big) = bd\operatorname{cov}(X,Y).\end{aligned}$$

Další tvrzení o rozptylu jsou už vcelku snadným důsledkem:

$$\begin{aligned}\operatorname{var}(X+Y) &= \mathrm{E}\big((X+Y)-\mathrm{E}(X+Y)\big)^2 = \\ &= \mathrm{E}\big((X-\mathrm{E}\,X)+(Y-\mathrm{E}\,Y)\big)^2 = \\ &= \mathrm{E}(X-\mathrm{E}\,X)^2 + 2\,\mathrm{E}(X-\mathrm{E}\,X)(Y-\mathrm{E}\,Y) + \\ &\quad + \mathrm{E}(Y-\mathrm{E}\,Y)^2 = \\ &= \operatorname{var} X + 2\operatorname{cov}(X,Y) + \operatorname{var} Y.\end{aligned}$$

Pokud jsou navíc X a Y nezávislé, jsou jistě nezávislé i náhodné veličiny $X-\mathrm{E}\,X$ a $Y-\mathrm{E}\,Y$. Pak je ovšem platí $\mathrm{E}(XY) = \mathrm{E}\,X\,\mathrm{E}\,Y$ a je tedy přímo z definice jejich kovariance nulová. $\square$

Přímo z definice také vidíme, že

$$\operatorname{var}(X) = \operatorname{cov}(X,X)$$

Y^2, XY:

$$\begin{aligned} E(X) &= E(Y) &&= 0\cdot\frac{1}{4}+1\cdot\frac{1}{2}+2\cdot\frac{1}{4}=1\\ E(X^2) &= E(Y^2) &&= 0\cdot\frac{1}{4}+1\cdot\frac{1}{2}+4\cdot\frac{1}{4}=\frac{3}{2}\\ E(XY) & &&= 0\cdot\frac{3}{8}+1\cdot\frac{1}{4}+2\cdot\frac{1}{4}+4\cdot\frac{1}{8}=\frac{5}{4} \end{aligned}$$

Máme tedy

$$\begin{aligned} \sigma^2(X) &= \varsigma^2(Y) &&= E(X^2)-[E(X)]^2=\frac{1}{2}\\ \operatorname{cov}(X,Y) & &&= E(XY)-E(X)E(Y)=\frac{1}{4} \end{aligned}$$

Celkem

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma(X)\cdot\sigma(Y)}=\frac{1}{2}$$

□

9.55. Nechť náhodné veličiny U, V mají diskrétní rozdělení určené následující tabulkou (U může nabývat hodnot 1, 2, veličina V potom hodnot 1, 2 a 3):

	V		
U	1	2	3
1	0,1	0,2	0,3
2	0,2	0,1	0,1

Najděte marginální rozdělení obou náhodných veličin, jejich střední hodnoty, rozptyly a korelační koeficient. ○

9.56. Určete střední hodnotu a rozptyl náhodné veličiny X^2, kde X je náhodná veličina s rovnoměrným rozdělením pravděpodobnosti na intervalu $\langle -1, 1\rangle$. ○

9.57. Dvakrát hodíme šestibokou kostkou. Určete korelační koeficient veličiny X udávající počet padlých sudých čísel a veličiny Y udávající počet padlých lichých čísel. ○

9.58. Nechť náhodné veličiny U, V mají rozdělení pravděpodobnosti určené následující tabulkou (U může nabývat hodnot 1, 2, veličina V potom hodnot 1, 2 a 3):

	V		
U	1	2	3
1	0,1	0,1	0,4
2	0,2	0,1	0,1

Najděte marginální rozdělení obou náhodných veličin, jejich střední hodnoty, rozptyly a korelační koeficient. ○

9.37. Korelace náhodných veličin. Předchozí věta nám říká, že

kovariance je symetrickou bilineární formou na reálném vektorovém prostoru náhodných veličin s rozptylem. Rozptyl je pak příslušnou kvadratickou formou a kovarianci lze pak spočíst z rozptylu jednotlivých veličin a jejich součtu, tak jak jsme to viděli v lineární algebře.

Kovariance tedy do jisté míry vypovídá o závislosti dvou náhodných veličin. Hovoříme o *korelaci veličin* a v obdobě ke směrodatné odchylce zavádíme následující pojem

KORELAČNÍ KOEFICIENT

Korelačním koeficientem náhodných veličin X a Y, které mají konečný nenulový rozptyl, rozumíme hodnotu

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}X}\sqrt{\operatorname{var}Y}}.$$

Jak je vidět z věty 9.36, korelační koeficient veličin je kovariance normovaných veličin $\frac{1}{\sqrt{\operatorname{var}X}}(X-\operatorname{E}X)$ a $\frac{1}{\sqrt{\operatorname{var}Y}}(Y-\operatorname{E}Y)$).

Okamžitě je také vidět platnost následujících vztahů, kde a, b, c, d, $bd \neq 0$, jsou reálné konstanty a X, Y jsou náhodné veličiny s nenulovým konečným rozptylem,

$$\begin{aligned} \rho_{a+bX,c+dY} &= \operatorname{sgn}(bd)\rho_{X,Y}\\ \rho_{X,X} &= 1. \end{aligned}$$

Navíc je jistě $\rho_{X,Y}=0$, pokud jsou náhodné veličiny X a Y nezávislé.

Všimněme si, že když má náhodná veličina X nulový rozptyl, pak přímo z definice vidíme, že musí nabývat hodnotu $\operatorname{E}X$ s pravděpodobností 1. Skutečně, kdyby padla hodnota X do nějakého intervalu I neobsahujícího $\operatorname{E}X$ s pravděpodobností $p\neq 0$, pak by musel být výraz $\operatorname{var}X=\operatorname{E}(X-\operatorname{E}X)^2$ kladný. Stochasticky se tedy veličiny s nulovým rozptylem chovají jako konstanty.

Kdyby byla kovariance pozitivně definitní symetrická bilineární forma, Cauchyova-Schwarzova nerovnost (viz 3.25) by okamžitě dala nerovnost

$$|\rho_{X,Y}|\leq 1 \tag{9.17}$$

V následující větě říkáme více. Ukazuje totiž, že korelace nebo antikorelace veličin X a Y říká, že jsou tyto veličiny v nějakém afinním vztahu $Y=kX+c$, přičemž znaménko k odpovídá znaménku v $\rho_{X,Y}=\pm 1$. Naopak, nulový korelační koeficient vypovídá o skutečnosti, že případnou závislost veličin vůbec nejde přiblížit pomocí takového afinního vztahu (a nemusí proto nutně jít o nezávislé veličiny).

Věta. *Je-li korelační koeficient definován, pak platí* $|\rho_{X,Y}|\leq 1$. *Rovnost přitom nastává pouze tehdy, když existují konstanty* k, c *takové, že* $P(Y=kX+c)=1$.

DŮKAZ. Protože je rozptyl vždy nezáporný, odhadneme kvadratický výraz

$$0\leq \operatorname{var}\left(\frac{Y-\operatorname{E}Y}{\sqrt{\operatorname{var}Y}}+t\,\frac{X-\operatorname{E}X}{\sqrt{\operatorname{var}X}}\right)=1+2t\rho_{X,Y}+t^2.$$

Kvadratický výraz napravo tedy jistě nemá dva reálné různé kořeny a proto musí být jeho diskriminant nekladný, tj. $4(\rho_{X,Y})^2-4\leq 0$. Odtud již dostáváme dokazovanou nerovnost a také vidíme, že rovnost nastává pouze pro $\rho_{X,Y}=\pm 1$. Pak ovšem pro jediný dvojnásobný kořen t_0 má příslušná veličina nulový rozptyl a má tedy s pravděpodobností jedna vhodnou konstantní hodnotu. □

G. Transformace náhodných veličin

Uvažme spojitou funkci náhodné veličiny $Y = \psi(X)$. Za předpokladu, že transformace ψ je rostoucí (analogicky klesající) funkce, dostáváme pro příslušnou distribuční funkci vztah

$$F_Y(y) = P(Y \le y) = P(\psi(X) \le y) = \\ = P(X \le \psi^{-1}(y)) = F_X(\psi^{-1}(y)),$$

kde F_X je distribuční funkce X. Odkud pro hustotu transformované náhodné veličiny Y

$$f_Y(y) = \frac{dF_Y(y)}{dy} = f_X(\psi^{-1}(y)) \left| \frac{d\psi^{-1}(y)}{dy} \right|.$$

Podle pravidla pro transformaci souřadnic v integrálu pak můžeme střední hodnotu Y spočítat jako

$$\mathrm{E}\, Y = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_{-\infty}^{\infty} \psi(x) f_X(x) dx$$

a podobně pro rozptyl Y.

9.59. Náhodná veličina X má hustotu $f(x)$. Určete hustotu náhodné veličiny Y tvaru

i) $Y = e^X, x \ge 0$,
ii) $Y = \sqrt{X}, x > 0$,
iii) $Y = \ln X, x > 0$,
iv) $Y = \frac{1}{X}, x > 0$.

Řešení. Přímým aplikováním formule pro hustotu transformované náhodné veličiny dostaneme a) $f_Y(y) = f(\ln y)\frac{1}{y}$, b) $f_Y(y) = 2f(y^2)y$, c) $f_Y(y) = f(e^y)e^y$, d) $f_Y(y) = f(1/y)\frac{1}{y^2}$. □

9.60. Náhodná veličina X má rovnoměrné rozdělení pravděpodobnosti na intervalu $(-\frac{\pi}{2}, \frac{\pi}{2})$. Určete jeho hustotu a hustotu transformovaných veličin $Y = \sin X$, $Z = \mathrm{tg}\, X$.

Řešení. Protože délka intervalu, na kterém je náhodná veličina X nenulová je π, je její hustota rovna $f_X(x) = \frac{1}{\pi}$ pro $x \in (-\frac{\pi}{2}, \frac{\pi}{2})$ a nula jinde. Ze vztahu pro hustotu transformované náhodné veličiny a podle vzorce pro derivaci elementárních funkcí pak máme

$$f_Y(y) = f_X(\arcsin(y)) \arcsin'(y) = \frac{1}{\pi\sqrt{1-y^2}}$$

a

$$f_Z(y) = f_X(\mathrm{arctg}(z)) \,\mathrm{arctg}'(y) = \frac{1}{\pi(1+y^2)}.$$ □

9.61. Náhodná veličina X má hustotu rovnu $\cos x$ pro $x \in (0, \frac{\pi}{2})$ a nulovou jinde. Určete hustotu náhodné veličiny $Y = X^2$ a vypočtěte $\mathrm{E}\, Y$, $\mathrm{var}\, Y$.

9.38. Varianční matice. Dostáváme se konečně k variabilitě hodnot náhodného vektoru. Nabízí se uvažovat kovariance všech dvojic komponent. Následující definice a věta ukazují, že skutečně dostaneme analogii rozptylu pro vektory, včetně chování rozptylu při afinních transformacích náhodných veličin.

VARIANČNÍ MATICE

Uvažme náhodný vektor $X = (X_1, \ldots, X_n)^T$ jehož všechny komponenty mají konečný rozptyl. *Varianční matici* náhodného vektoru X definujeme pomocí střední hodnoty předpisem (vektor X je sloupec náhodných veličin)

$$\mathrm{var}\, X = \mathrm{E}(X - \mathrm{E}\, X)(X - \mathrm{E}\, X)^T.$$

Použitím definice střední hodnoty vektoru a přímým rozepsáním násobení matic po složkách ověříme, že varianční matice je symetrická matice

$$\mathrm{var}\, X = \begin{pmatrix} \mathrm{var}\, X_1 & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}\, X_2 & \cdots & \mathrm{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_n, X_1) & \mathrm{cov}(X_n, X_2) & \cdots & \mathrm{var}\, X_n \end{pmatrix}$$

Věta. *Uvažujme náhodný vektor $X = (X_1, \ldots, X_n)^T$, jehož všechny komponenty mají konečný rozptyl, Uvažme dále jeho transformovanou vektorovou náhodnou veličinu $Y = BX + c$, kde B je matice reálných konstant typu $m \times n$ a c je vektor konstant v $\mathbb{R}^m$. Potom*

$$\mathrm{var}(Y) = \mathrm{var}(BX + c) = B(\mathrm{var}\, X)B^T.$$

DŮKAZ. Stačí provést přímý výpočet a využít přitom vlastnosti střední hodnoty

$$\mathrm{var}(Y) = \mathrm{E}\big((BX + c) - \mathrm{E}(BX + c)\big)\big((BX + c) - \\ - \mathrm{E}(BX + c)\big)^T = \\ = \mathrm{E}(B(X - \mathrm{E}\, X))(B(X - \mathrm{E}\, X))^T = \\ = B\,\mathrm{E}(X - \mathrm{E}\, X)(X - \mathrm{E}\, X)^T B^T = \\ = B(\mathrm{var}\, X)B^T.$$

□

Stejně jako u rozptylu skalární náhodné veličiny tedy vidíme, že konstantní část transformace nemá vliv, zatímco vůči lineární části transformace se varianční matice chová jako matice kvadratické formy.

9.39. Momenty a momentová funkce. Střední hodnota a rozptyl odráží chování střední hodnoty samotné veličiny X a jejího kvadrátu. V popisné statistice jsme také zkoumali tzv. šikmost rozložení dat a je přirozené zkoumat variabilitu náhodných veličin pomocí vyšších mocnin dané náhodné veličiny X.

Charakteristiku $E(X^k)$ nazýváme *k-tým momentem*, charakteristiku $\mu_k = E\left((X - EX)^k\right)$ pak *k-tým centrálním momentem* náhodné veličiny X. Užitečný bývá také tzv. *k-tý absolutní moment* zadaný předpisem $\mathrm{E}\,|X|^k$.

Přímo z definice je tedy pro spojitou veličinu X

$$\mathrm{E}\, X^k = \int_{-\infty}^{\infty} x^k f_X(x)\, dx$$

Řešení. Podle vzorce pro hustotu transformované náhodné veličiny je

$$f_Y(y) = f_X(\sqrt{y})(\sqrt{y})' = \frac{1}{2\sqrt{y}}\cos x.$$

Střední hodnotu a rozptyl Y je jednodušší počítat přímo z hustoty náhodné veličiny X. Platí $\mathrm{E}\,Y = \int_{-\infty}^{\infty} x^2 f_X(x)dx$ a proto

$$\mathrm{E}\,Y = \int_0^{\frac{\pi}{2}} x^2 \cos x dx = \left[x^2 \sin x + 2x\cos x - 2\sin x\right]_0^{\frac{\pi}{2}} = \frac{\pi^2 - 8}{4}.$$

Integrál jsme spočítali metodou per partes. Stejnou metodou spočítáme

$$\mathrm{E}(Y^2) = \int_0^{\frac{\pi}{2}} x^4 \cos x dx = \\ = \left[(x^4 - 12x^2 + 24)\sin x + 4(x^3 - 6x)\cos x\right]_0^{\frac{\pi}{2}}.$$

Odtud máme $\mathrm{E}(Y^2) = (\frac{\pi}{2})^4 - 12(\frac{\pi}{2})^2 + 24$, a proto $\operatorname{var} Y = \frac{\pi^4}{16} - 3\pi^2 + 24 - \frac{\pi^4 - 16\pi^2 + 64}{16} = 20 - 2\pi^2$. □

9.62. Nechť X je náhodná veličina, která nabývá hodnoty 0 s pravděpodobností $\frac{1}{2}$ a hodnoty 1 též s pravděpodobností $\frac{1}{2}$. Podobně nechť Y je náhodná veličina, která nabývá hodnoty -1 a 1 s pravděpodobnostmi $\frac{1}{2}$. Ukažte, že náhodné veličiny X a $Z = XY$ jsou nekorelované, ale závislé. Udejte příklad dvou spojitých náhodných veličin, které mají tuto vlastnost.

Řešení. Nejprve spočítáme střední hodnoty našich náhodných veličin $\mathrm{E}\,X = 0\cdot\frac{1}{2} + 1\cdot\frac{1}{2} = \frac{1}{2}$, $\mathrm{E}\,Z = \mathrm{E}(XY) = 0\cdot\frac{1}{2} + (-1)\cdot\frac{1}{4} + 1\cdot\frac{1}{4} = 0$. Pro střední hodnotu jejich součinu máme $\mathrm{E}(XZ) = \mathrm{E}(X^2Y) = 1\cdot\frac{1}{4} + (-1)\cdot\frac{1}{4} = 0$. Podle věty 9.36 je pak kovariance rovna $\operatorname{cov}(X,Z) = 0 - \frac{1}{2}\cdot 0 = 0$. Veličiny X a Y jsou tedy nekorelované. Zároveň je podmíněná pravděpodobnost $P(Z = 1|X = 0)$ zřejmě nulová, tj. i $P(Z = 1, X = 0) = 0$, a přitom $P(Z = 1) = \frac{1}{4}$ a $P(X = 0) = \frac{1}{2}$, tedy $P(Z = 1)\cdot P(X = 0) = \frac{1}{8} \neq 0$. Vidíme, že $P(Z = 1)\cdot P(X = 0) \neq P(Z = 1, X = 0)$, což znamená, že X a Z jsou závislé.

Z příslušných definic lze lehce ověřit, že příkladem spojitých nekorelovaných závislých náhodných veličin jsou X a $Y = X^2$, kde X je libovolně rozložená náhodná veličina, která má nulovou střední hodnotu, konečný druhý moment a nulový třetí moment. □

H. Nerovnosti a limitní věty

Markovova nerovnost dává hrubý odhad nezáporné náhodné veličiny v případě, že neznáme nic jiného, než její střední hodnotu. Konkrétně říká, že pro každou nezápornou náhodnou veličinu X a pro libovolné $a > 0$ platí $P(X \geq a) \leq \frac{\mathrm{E}\,X}{a}$.

a obdobně víme, že pro diskrétní veličiny X s pravděpodobností soustředěnou do hodnot x_i bude

$$\mathrm{E}\,X^k = \sum_i x_i^k f_X(x_i).$$

Uvidíme, že bude pro výpočty velice výhodné umět pracovat s mocninnou řadou, ve které momenty budou vystupovat coby koeficienty. Protože víme, že koeficienty Taylorovy řady funkce $M(t)$ v bodě $t = 0$ dostaneme pomocí diferencování, můžeme vcelku snadno uhádnout správnou volbu takové funkce:

MOMENTOVÁ VYTVOŘUJÍCÍ FUNKCE

Pro náhodnou veličinu X uvažme funkci $M_X(t) : \mathbb{R} \to \mathbb{R}$ definovanou předpisem

$$M_X(t) = \mathrm{E}\,e^{tX} = \begin{cases} \sum_i \mathrm{e}^{tx_i} f_X(x_i) & \text{pro diskrétní } X \\ \int_{-\infty}^{\infty} \mathrm{e}^{tx} f_X(x)\,dx & \text{pro spojitou } X. \end{cases}$$

Pokud tato střední hodnota existuje, hovoříme o *momentové vytvořující funkci* náhodné veličiny X.

Je zjevné, že tato funkce $M_X(t)$ je vždy analytickou funkcí v případě diskrétních náhodných veličin s konečně mnoha hodnotami x_i.

Věta. *Nechť X je náhodná veličina pro kterou na intervalu $(-a, a)$ existuje její analytická momentová vytvořující funkce. Pak na tomto intervalu je $M_X(t)$ dána absolutně konvergující řadou*

$$M_X(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!}\mathrm{E}\,X^k.$$

DŮKAZ. Ověření tvrzení věty je jednoduchým cvičením na techniky diferenciálního a integrálního počtu. V případě diskrétní veličiny, jde buď o konečné součty nebo o počítání s absolutně a stejnoměrně konvergentními řadami, resp. v případě spojitých veličin jde o absolutně konvergující integrály. Můžeme proto prohodit limitní proces s derivováním a protože $\frac{d}{dt}\mathrm{e}^{tx} = x\,\mathrm{e}^{tx}$, dostáváme okamžitě vztah

$$\frac{d^k}{dt^k}M_X(t) = \mathrm{E}\,X^k$$

a odtud je tvrzení věty zřejmé. □

Ve skutečnosti lze ukázat, že předpoklady věty jsou splněny, kdykoliv platí současně $M_X(-a) < \infty$ a $M_X(a) < \infty$ a navíc lze dokázat, že platí-li v takovém případě rovnost momentových funkcí $M_X(t) = M_Y(t)$ na nějakém netriviálním intervalu, pak mají tyto náhodné veličiny X a Y také stejné distribuční funkce. Momentová funkce tedy poskytuje za těchto podmínek úplnou charakterizaci náhodné veličiny.

9.40. Vlastnosti momentové funkce. Díky vlastnostem exponenciální funkce lze očekávat, že snadno spočteme, jak se chová momentová vytvořující funkce při afinních transformacích náhodných veličin a při součtech nezávislých náhodných veličin.

Lemma. *Nechť $a, b \in \mathbb{R}$ a X, Y jsou nezávislé náhodné veličiny s momentovými vytvořujícími funkcemi $M_X(t)$ a $M_Y(t)$. Potom*

9.63. Mějme nezápornou náhodnou veličinu X se střední hodnotou μ. Bez dalších informací o rozdělení X odhadněte $P(X > 3\mu)$. Vypočtěte $P(X > 3\mu)$ víte-li, že $X \sim \mathrm{Ex}(\frac{1}{\mu})$.

Řešení. Pokud nezáporná náhodná veličina X nenabývá pouze nulovou hodnotu, pak je její střední hodnota μ kladná. Proto můžeme danou pravděpodobnost zhruba odhadnout pomocí Markovovy nerovnosti

$$P(X \geq 3\mu) \leq \frac{\mu}{3\mu} = \frac{1}{3}.$$

Pokud víme, že $X \sim \mathrm{Ex}(\frac{1}{\mu})$, pak

$$P(X > 3\mu) = 1 - P(X \leq 3\mu) = 1 - F(3\mu),$$

kde F je distribuční funkce exponenciálního rozdělení. Ta je podle definice

$$F(x) = \int_0^x \frac{1}{\mu} e^{-\frac{t}{\mu}} dt = \left[-e^{-\frac{t}{\mu}}\right]_0^x = 1 - e^{-\frac{x}{\mu}}$$

a proto $P(X > 3\mu) = \frac{1}{e^3}$. □

9.64. Průměrná rychlost větru je na určitém místě 20 km/hod.

- Bez ohledu na rozdělení rychlosti větru jako náhodné veličiny odhadněte pravděpodobnost, že při jednom pozorování rychlost větru nepřesáhne 60 km/h.
- Určete interval, v němž se bude rychlost větru nacházet s pravděpodobností alespoň 0,9, víte-li navíc, že směrodatná odchylka $\sigma = 1$ km/hod.

Řešení. Označme náhodnou veličinu udávající rychlost větru X. V prvním případě můžeme použít pouze hrubý odhad pomocí Markovovy nerovnosti

$$P(X \leq 60) = 1 - P(X \geq 60) \geq 1 - \frac{20}{60} = \frac{2}{3}.$$

V druhém případě známe rozptyl (resp. směrodatnou odchylku) rychlosti větru, a proto k určení daného intervalu můžeme použít Čebyševovu nerovnost 9.35

$$0,9 \leq P(|X - 20| < x) = 1 - P(|X - 20| \geq x) \leq 1 - \frac{1}{x^2}.$$

Odtud $x \geq \sqrt{10} \approx 3,2$. Hledaný interval je tedy (16,8 km/hod, 23,2 km/hod). □

9.65. Ke každému jogurtu běžné značky je náhodně (rovnoměrně) přibalen obrázek některého z 26 hokejových mistrů světa. Kolik jogurtů si fanynka Věrka musí koupit, aby s pravděpodobností 0,95 získala alespoň 5 kartiček Jaromíra Jágra?

Řešení. Označíme-li náhodnou veličinu udávající počet získaných kartiček Jágra X, je zřejmě $X \sim \mathrm{Bi}(n, \frac{1}{26})$, kde n je celkový počet koupených jogurtů. Hledáme takovou hodnotu tohoto čísla, aby

mají náhodné veličiny $V = a + bX$ a $W = X + Y$ momentové vytvořující funkce

$$M_{a+bX}(t) = \mathrm{e}^{at} M_X(bt)$$
$$M_{X+Y}(t) = M_X(t) M_Y(t)$$

Důkaz. První vztah spočteme přímo z definice

$$M_V(t) = \mathrm{E}\,\mathrm{e}^{(a+bX)t} = \mathrm{E}\,\mathrm{e}^{at}\,\mathrm{e}^{(bt)X} = \mathrm{e}^{at} M_X(bt).$$

U druhého využijeme skutečnost, že střední hodnota součinu nezávislých veličin je součinem jejich středních hodnot.

$$M_W(t) = \mathrm{E}\,\mathrm{e}^{t(X+Y)} = \mathrm{E}\,\mathrm{e}^{tX}\,\mathrm{e}^{tY} = \mathrm{E}\,\mathrm{e}^{tX}\,\mathrm{E}\,\mathrm{e}^{tY} = M_X(t) M_Y(t).$$

□

Pro ilustraci si spočtěme přímo z definice momentovou funkci náhodné veličiny X s normálním rozložením $\mathrm{N}(\mu, \sigma)$ a náhodné veličiny X s binomiálním rozložením $\mathrm{Bi}(n, p)$. Začneme s veličinou $Z \sim \mathrm{N}(0, 1)$

$$\begin{aligned} M_Z(t) &= \int_{-\infty}^{\infty} \mathrm{e}^{tx} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x = \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(x^2 - 2tx + t^2 - t^2)}\,\mathrm{d}x = \\ &= \mathrm{e}^{\frac{t^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-t)^2}{2}}\,\mathrm{d}x = \\ &= \mathrm{e}^{\frac{t^2}{2}}, \end{aligned}$$

kde jsme využili při výpočtu skutečnost, že v předposledním výrazu integrujeme pro každé pevné t hustotu rozdělení spojité náhodné veličiny, proto je tento integrál roven jedné.

Jde tedy o případ všude analytické funkce a zejména existují momenty všech řádů. Přímým dosazením $\frac{1}{2}t^2$ do mocninné řady pro exponenciálu je všechny okamžitě spočteme:

$$\begin{aligned} M_Z(t) &= \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{t^2}{2}\right)^k = \sum_{k=0}^{\infty} \frac{1}{k! 2^k} t^{2k} = \\ &= 1 + 0\,t + \frac{1}{2}\,t^2 + 0\,t^3 + \frac{3}{4!}\,t^4 + \ldots \end{aligned}$$

Zejména tedy znovu vidíme, že střední hodnota Z je skutečně $\mathrm{E}\,Z = 0$ a její rozptyl je $\mathrm{var}\,Z = \mathrm{E}\,Z^2 - (\mathrm{E}\,Z)^2 = 1$.

Dosazením do vztahu pro momentovou vytvořující funkci $M_{\mu+\sigma Z}$ dostaneme pro $X \sim \mathrm{N}(\mu, \sigma)$

$$M_X(t) = \mathrm{e}^{\mu t}\,\mathrm{e}^{\frac{\sigma^2 t^2}{2}}$$

a odtud také okamžitě vidíme, že součet nezávislých normálních rozdělení $X \sim \mathrm{N}(\mu, \sigma)$ a $Y \sim \mathrm{N}(\mu', \sigma')$ má opět normální rozdělení $X + Y \sim \mathrm{N}(\mu + \mu', \sigma + \sigma')$.

Podobně pro veličinu $X \sim \mathrm{Bi}(n, p)$ spočteme snadno

$$\begin{aligned} M_X(t) = \mathrm{E}\,\mathrm{e}^{tX} &= \sum_{k=0}^{n} (p\,\mathrm{e}^t)^k \binom{n}{k} (1-p)^{n-k} = \\ &= (p\,\mathrm{e}^t + (1-p))^n = (p(\mathrm{e}^t - 1) + 1)^n = \\ &= 1 + npt + (\binom{n}{2} p^2 + n\frac{p}{2}) t^2 + \ldots. \end{aligned}$$

$P(X \geq 5) = 0{,}95$, tj. $F_X(4) = P(X \leq 4) = 0{,}05$. Abychom ji mohli určit, aproximujeme binomické rozdělení podle Moivreovy-Laplaceovy věty normálním rozdělením (předpokládáme hodnota n bude velké, a proto chyba aproximace bude malá). Podle ‖F‖ má X střední hodnotu $\mathrm{E}\,X = \frac{n}{26}$ a rozptyl $\mathrm{var}\,X = \frac{25n}{26^2}$. Označíme-li tedy Z standardizovanou veličinu, pak danou podmínku můžeme ekvivalentně přepsat

$$0{,}05 = P(X \leq 4) = P\left(Z \leq \frac{4 - \frac{n}{26}}{\frac{5\sqrt{n}}{26}}\right) = F_Z\left(\frac{104 - n}{5\sqrt{n}}\right),$$

kde $F_Z \approx \Phi$ je podle aproximačního předpokladu distribuční funkce normálního rozdělení N(0, 1). Protože určitě $n > 104$, tak využitím $\Phi(-x) = 1 - \Phi(x)$ předchozí rovnice dává $n - 104 = \Phi^{-1}(0{,}95) \cdot 5\sqrt{n}$. Kvantil vystupující v této rovnici má podle tabulek hodnotu $z(0{,}95) = 1{,}65$. Vyřešením této kvadratické rovnice pak obdržíme $n \doteq 228{,}8$. Věrka tedy musí koupit aspoň 229 jogurtů. □

9.66. Určete pravděpodobnost, že při 1200 hodech kostkou padne šestka alespoň 150 krát a nejvýše 250 krát pomocí Čebyševovy nerovnosti a pak pomocí Moivreovy-Laplaceovy věty.

Řešení. Označíme-li náhodnou veličinu udávající počet šestek X, pak je zjevně $X \sim \mathrm{Bi}(1200, \frac{1}{6})$. Podle ‖F‖ je tedy $\mathrm{E}\,X = 1200 \cdot \frac{1}{6} = 200$ a $\mathrm{var}\,X = 200(1 - \frac{1}{6}) = \frac{500}{3}$. Podmínka na počet šestek má ze zadání tvar $150 \leq X \leq 250$, což lze zapsat také jako $|X - 200| \leq 50$. Použitím Čebyševovy nerovnosti 9.35 pak

$$P(|X - 200| \leq 50) = 1 - P(|X - 200| \geq 51) \geq 1 - \frac{500}{3 \cdot 51^2} \approx 0{,}94.$$

(2) Přesná hodnota hledané pravděpodobnosti je zřejmě dána výrazem

$$P(150 \leq X \leq 250) = F_X(250) - F_X(150),$$

kde F_X je distribuční funkce binomického rozdělení. Z definice tedy

$$P(150 \leq X \leq 250) = \sum_{k=150}^{250} \binom{1200}{x} \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{1200-x}.$$

Tento výraz je obtížně vyčíslitelný, a proto k jeho odhadu využijeme Moivreovu-Laplaceovu větu. Nahradíme-li X standardizovanou náhodnou veličinou

$$Z = \frac{\sqrt{3}(X - 200)}{10\sqrt{5}},$$

pak podle 9.42 je $Z \sim N(0, 1)$, tj. $F_Z \approx \Phi$, a tedy

$$P(150 \leq X \leq 250) = P\left(\tfrac{\sqrt{3}(150-200)}{10\sqrt{5}} \leq Z \leq \tfrac{\sqrt{3}(250-200)}{10\sqrt{5}}\right) \approx$$
$$\approx \Phi(\sqrt{15}) - \Phi(-\sqrt{15}) = 2\Phi(\sqrt{15}) - 1.$$

Z tabulek $\Phi(\sqrt{15}) \approx 0{,}99994$, a proto je hledaná pravděpodobnost asi 99,988%. □

Samozřejmě jsme mohli totéž spočíst ještě snadněji s využitím posledního lemmatu, protože je X součtem n nezávislých veličin $Y \sim A(p)$ s alternativním rozdělením. Je tedy nutně

$$\mathrm{E}\,\mathrm{e}^{tX} = (\mathrm{E}\,\mathrm{e}^{tY})^n = (p\,\mathrm{e}^t + (1 - p))^n.$$

Opět odtud hned vidíme, že všechny momenty veličiny Y jsou rovny p. Proto $\mathrm{E}\,Y = p$, zatímco $\mathrm{var}\,Y = p(1 - p)$. Z momentové funkce $M_X(t)$ odečteme snadno $\mathrm{E}\,X = np$ a $\mathrm{var}\,X = \mathrm{E}\,X^2 - (\mathrm{E}\,X)^2 = np(1 - p)$.

Všimněme si, že náhodná veličina vzniklá jako součet n nezávislých náhodných veličin Y_i se stejným rozložením se samozřejmě stochasticky chová zásadně odlišně od násobku nY.

9.41. Šikmost a špičatost. Protože je třetí centrální moment dán pomocí třetích mocnin odchylek od střední hodnoty, bude do jisté míry vyjadřovat, jak moc nejsou hodnoty náhodné veličiny rozprostřeny symetricky kolem střední hodnoty. To jsme v popisné statistice sledovali pomocí koeficientu šikmosti. U náhodných veličin se používá se v podobě charakteristiky

$$\gamma_1 = \frac{\mathrm{E}(X - \mathrm{E}\,X)^3}{(\sqrt{\mathrm{var}\,X})^3}$$

a říkáme jí *koeficient šikmosti náhodné veličiny* X.

Další běžně užívanou charakteristikou je *koeficient špičatosti* náhodné veličiny X, který definujeme předpisem

$$\gamma_2 = \frac{\mathrm{E}(X - \mathrm{E}\,X)^4}{(\mathrm{var}\,X)^2} - 3.$$

Viděli jsme, že u normovaného normálního rozdělení je třetí centrální moment nulový a čtvrtý je roven 3. Zvolené normování koeficientu špičatosti je voleno tak, aby jeho hodnota pro normované normální rozdělení byla nulová. Pro obecné rozložení pak špičatost dává srovnání s normálním rozdělením.

V praxi se však můžeme setkat i s jinými normováními koeficientů šikmosti a špičatosti.

9.42. Centrální limitní věta. Nyní se konečně dostáváme ke klíčovému nástroji, který propojuje pravděpodobnost a statistiku. Technicky se bude zdát, že jde o vcelku jednoduchou manipulaci s momentovými vytvořujícími funkcemi. Historicky však byly daleko dříve a jinak dokázány mnohé speciální případy, které samy o sobě mají velkou hodnotu, protože často podávají navíc odhady rychlosti konvergence, a ty jsou pochopitelně pro praktické využití třeba.

Před formulací výsledku se nejprve zastavme u zobecnění Bernoulliovy věty o binomickém rozdělení na konci odstavce 9.35. Náhodné veličiny $\frac{1}{n}X_n$, kde $X_n \sim \mathrm{Bi}(n, p)$ můžeme považovat za aritmetický průměr součtu n nezávislých veličin s rozdělením $\mathrm{A}(p)$ a samotné Bernoulliho tvrzení pak říká, že tyto průměry konvergují k hodnotě p s pravděpodobností 1. Toto tvrzení platí zcela obecně takto:

Lemma. *Uvažme posloupnost po dvou nekorelovaných náhodných veličin $X_1, X_2, \ldots$, které mají všechny společnou konečnou střední hodnotu $\mathrm{E}\,X_i = \mu$. Předpokládejme navíc, že tyto veličiny mají konečné rozptyly omezené konstantou $\mathrm{var}\,X_i \leq C$. Potom pro libovolné $\varepsilon > 0$ platí*

$$\lim_{n\to\infty} P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\Big| < \varepsilon\Big) = 1.$$

9.67. Na fakultě informatiky je 10% studentů s prospěchem do 1,2. Jak velkou skupinu je třeba vybrat, aby s pravděpodobností 0,95 v ní bylo 8-12% studentů s prospěchem do 1,2? Úlohu řešte napřed pomocí Čebyševovy a potom pomocí Moivre-Laplaceovy věty.

Řešení. Označme jako X náhodnou veličinu udávající počet studentů s prospěchem do 1,2 z n vybraných studentů. Při výběru jednotlivého studenta vyberu takového s pravděpodobností 10%, a proto při nezávislém výběru n studentů je $X \sim \mathrm{Bi}(n, \frac{1}{10})$. Podle ‖F‖ je $\mathrm{E}\,X = 0{,}1n$ a $\mathrm{var}\,X = 0{,}09n$. Pro hledanou pravděpodobnost pak podle Čebyševovy nerovnosti 9.35 platí

$$P(|X - 0{,}1n| \le 0{,}02n) = 1 - P(|X - 0{,}1n| \ge 0{,}02n) \ge$$
$$\ge 1 - \frac{0{,}1 \cdot 0{,}9n}{(0{,}02n)^2} = 1 - \frac{225}{n}.$$

Nerovnost $1 - \frac{225}{n} \ge 0{,}95$ a tedy i

$$P(|X - 0{,}1n| \le 0{,}02n) \ge 0{,}95.$$

je splněna pro $n \ge 4500$. Přesná hodnota pravděpodobnosti je dána pomocí distribuční funkce F_X binomického rozdělení

$$P(0{,}08n \le X \le 0{,}12n) = F_X(0{,}12n) - F_X(0{,}08n).$$

Podle Moivreovy-Laplaceovy věty z 9.42 můžeme standardizovanou náhodnou veličinu $Z = \frac{10X-n}{3\sqrt{n}}$ aproximovat normovaným normálním rozložením, $F_Z \approx \Phi$, a proto

$$\begin{aligned} 0{,}95 &= P(0{,}08n \le X \le 0{,}12n) = P(-\frac{\sqrt{n}}{15} \le Z \le \frac{\sqrt{n}}{15}) \approx \\ &\approx \Phi(\frac{\sqrt{n}}{15}) - \Phi(-\frac{\sqrt{n}}{15}) = \\ &= 2\Phi(\frac{\sqrt{n}}{15}) - 1. \end{aligned}$$

Odtud $\sqrt{n} = 15z(0{,}975)$ a z tabulek dopočítáme $n \approx 864{,}4$. Vidíme tedy, že stačí vybrat 865 studentů. □

9.68. Pravděpodobnost, že zasazený strom se ujme, je 0,8. Jaká je pravděpodobnost, že z 500 zasazených stromů se jich ujme aspoň 380?

Řešení. Náhodná veličina X udávající počet stromů, které se ujaly, má binomické rozdělení $X \sim \mathrm{Bi}(500, \frac{4}{5})$. Podle ‖F‖ je $\mathrm{E}\,X = 400$ a $\mathrm{var}\,X = 80$. Standardizovaná náhodná veličina je tedy $Z = \frac{X-400}{\sqrt{80}}$. Podle Moivreovy-Laplaceovy věty je $F_Z \approx \Phi$, a proto

$$\begin{aligned} P(X \ge 380) &= P(Z \ge \frac{380 - 400}{\sqrt{80}}) \approx 1 - \Phi(-\frac{\sqrt{20}}{2}) = \\ &= \Phi(\frac{\sqrt{20}}{2}) \approx 0{,}987. \end{aligned}$$

□

Důkaz. Tvrzení ověříme pomocí Čebyševovy nerovnosti stejně, jako jsme postupovali v závěru odstavce 9.35. Spočteme

$$P\big(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| \ge \varepsilon\big) \le \frac{\mathrm{var}\big(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\big)}{\varepsilon^2} =$$
$$= \frac{\frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}\,X_i}{\varepsilon^2} \le \frac{C}{n\varepsilon^2}.$$

Je tedy pravděpodobnost zkoumaná v našem tvrzení odhadnuta zdola výrazem

$$P\big(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| < \varepsilon\big) \ge 1 - \frac{C}{n\varepsilon^2}$$

a lemma je dokázáno. □

Vidíme tedy, že k tomu, aby posloupnosti průměrů po dvou nekorelovaných veličin X_i s nulovou střední hodnotou konvergovaly (ve smyslu pravděpodobnosti) k nule, potřebujeme jen existenci a stejnoměrnou omezenost jejich rozptylů.

Náš další cíl bude ambicióznější. Budeme asymptotické chování posloupnosti náhodných veličin X_i porovnávat s normálním rozdělením. Chceme přitom uvažovat posloupnost nezávislých normovaných náhodných veličin se stejným rozdělením pravděpodobnosti, které však nemusí být ani normální ani binomické.

Předpokládáme tedy $\mathrm{E}\,X_i = 0$ a $\mathrm{var}\,X_i = 1$. Z technických důvodů dále předpokládejme, že existuje momentová vytvořující funkce $M_X(t)$ všech veličin X_i a že je také stejnoměrně omezený třetí absolutní moment $\mathrm{E}\,|X_i|^3 < C$.

Aritmetický průměr $\frac{1}{n}\sum_{i=1}^{n} X_i$ je samozřejmě náhodná veličina se střední hodnotou 0, její rozptyl je ale $\frac{n}{n^2} = \frac{1}{n}$. Uvažujme proto místo aritmetických průměrů raději náhodné veličiny

$$S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i,$$

které budou opět normované. Jejich momentové vytvořující funkce jsou (viz lemma 9.40)

$$M_{S_n}(t) = \mathrm{E}\,\mathrm{e}^{\frac{t}{\sqrt{n}}\sum_i X_i} = \Big(M_X(\frac{t}{\sqrt{n}})\Big)^n.$$

Vzhledem k předpokladu o normovanosti veličin X_i platí

$$M_X(\frac{t}{\sqrt{n}}) = 1 + 0\,\frac{t}{\sqrt{n}} + 1\,\frac{t^2}{2n} + \mathrm{o}(\frac{t^2}{n}),$$

kde opět píšeme $\mathrm{o}(G(n))$ pro výraz, který jde po podělení výrazem $G(n)$ v limitě pro $n \to \infty$ k nule, viz odstavec 6.17.

V limitě tedy můžeme psát (připomeňme, že třetí absolutní moment je ohraničený konstantou C)

$$\lim_{n\to\infty} M_{S_n}(t) = \lim_{n\to\infty}\Big(1 + \frac{t^2}{2n} + \mathrm{o}(\frac{1}{n})\Big)^n = \mathrm{e}^{\frac{t^2}{2}}.$$

To je ale právě momentová vytvořující funkce normálního rozdělení $Z \sim \mathrm{N}(0, 1)$, viz konec odstavce 9.38. Naše normované veličiny S_n tedy asymptoticky mají normované normální rozdělení. Tím jsme odvodili následující základní větu:

Věta (Centrální limitní věta)**.** *Uvažme posloupnost nezávislých náhodných veličin X_i, které mají společné střední hodnotu $\mathrm{E}\,X_i = \mu$*

9.69. Pomocí distribuční funkce standardního normálního rozdělení určete pravděpodobnost, že při 1600 hodech mincí bude rozdíl mezi počtem padlých hlav a orlů alespoň 82.

Řešení. Označíme-li jako X náhodnou veličinu udávající počet padlých hlav, tak X má binomické rozložení pravděpodobnosti $Bi(1600, 1/2)$ (se střední hodnotou 800 a směrodatnou odchylkou 20) a tudíž lze distribuční funkci veličiny $\frac{X-800}{20}$ lze pro dané velké $n = 1600$ podle Moivreovy-Laplaceovy věty velmi dobře odhadnout jako distribuční funkci Φ standardního normálního rozdělení. Hledaná pravděpodobnost je tedy

$$\begin{aligned} P &= 1 - P[759 \le X \le 841] \\ &= 1 - P\left[-2,05 \le \frac{X-800}{20} \le 2,05\right] \\ &\doteq 2\Phi(-2,05) \doteq 0,0404. \end{aligned}$$

□

9.70. Pomocí distribuční funkce standardního normálního rozdělení určete pravděpodobnost, že při 3600 hodech mincí bude rozdíl mezi počtem padlých hlav a orlů nejvýše 66.

Řešení. Označíme-li jako X náhodnou veličinu udávající počet padlých hlav, tak X má binomické rozložení pravděpodobnosti $Bi(3600, 1/2)$ (se střední hodnotou 1800 a směrodatnou odchylkou 30) a tudíž lze distribuční funkci veličiny $\frac{X-1800}{30}$ lze pro dané velké $n = 3600$ podle Moivreovy-Laplaceovy věty velmi dobře odhadnout jako distribuční funkci Φ standardního normálního rozdělení. Hledaná pravděpodobnost je tedy

$$\begin{aligned} P[1767 \le X \le 1833] &= P\left[-1,1 \le \frac{X-1800}{30} \le 1,1\right] \doteq \\ &\doteq \Phi(1,1) - \Phi(-1,1) \doteq 0,7498. \end{aligned}$$

□

9.71. Pravděpodobnost, že semeno vyklíčí, je 0,9. Kolik semen je třeba zasadit, aby s pravděpodobností aspoň 0,995 vyklíčilo cca 90% semen (což přesněji formulujeme se zpřesňujícím požadavkem, aby odchylka podílu vyklíčených semen od 0,9 nepřevýšila 0,034).

Řešení. Náhodná veličina X, udávající počet vyklíčených semen z n zasazených, má binomické rozdělení $X \sim \mathrm{Bi}(n, \frac{9}{10})$. Podle ‖F‖ je $\mathrm{E}\,X = 0,9n$ a $\mathrm{var}\,X = 0,09n$, a proto je standardizovaná veličina $Z = \frac{X-0,9n}{\sqrt{0,09n}}$. Podmínku ze zadání lze psát ve tvaru

$$\begin{aligned} P(|X - 0,9n| \le 0,034n) &= P\left(|Z| \le \frac{0,034n}{\sqrt{0,09n}}\right) = \\ &= P\left(|Z| \le \frac{0,34}{3}\sqrt{n}\right) \ge 0,995. \end{aligned}$$

a rozptyl $\mathrm{var}\,X_i = \sigma^2 > 0$ *a stejnoměrně omezený třetí absolutní moment* $\mathrm{E}\,|X_i|^3 < C$. *Pro rozdělení náhodné veličiny*

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)$$

platí v limitě vztah

$$\lim_{n\to\infty} P(S_n < x) = \Phi(x),$$

kde Φ *je distribuční funkce normovaného normálního rozdělení.*

Všimněme si, že v případě centrální limitní věty dostáváme jako výsledek asymptotické chování, které říká, že distribuční funkce jistých veličin se blíží k distribuční funkci normovaného normálního rozdělení. Takovému chování říkáme *konvergence podle distribuční funkce*. Je zřejmé, že tato konvergence je slabší než je konvergence podle pravděpodobnosti.

9.43. Moivreova-Laplaceova věta. Historicky asi první formulací centrální limitní věty byl případ veličin Y_n s binomickým rozdělením $\mathrm{Bi}(n, p)$. Ty můžeme chápat jako součet n nezávislých veličin X_i s alternativním rozdělením $A(p)$, $0 < p < 1$. Přitom jsme viděli, že tyto veličiny mají momentovou vytvořující funkci a $\mathrm{E}\,|X_i|^3 = p < 1$.

Centrální limitní věta v tomto případě tedy říká, že náhodné veličiny

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{X_i - p}{\sqrt{p(1-p)}}\right) = \frac{X - np}{\sqrt{np(1-p)}}$$

se asymptoticky chovají stejně jako normované normální rozdělení.

To také můžeme formulovat tak, že náhodná veličina $X \sim \mathrm{Bi}(n, p)$ se s rostoucím n chová jako veličina s normálním rozdělením $\mathrm{N}(np, np(1-p))$.

V praxi se považuje za vyhovující aproximace binomického rozdělení pomocí normálního, jestliže platí $np(1-p) > 9$.

Zkusme si výsledek ilustrovat na konkrétním příkladu. Řekněme, že chceme s chybou nejvýše 5% zjistit, kolik procent studentů má v oblibě danou přednášku. Počet osob majících přednášku v oblibě mezi n náhodně vybranými bude nejspíš mít charakter náhodné veličiny $X \sim \mathrm{Bi}(n, p)$. Dejme tomu, že přitom chceme, abychom dosáhli správného výsledku se spolehlivostí (tj. opět pravděpodobností) alespoň 90%. Chceme tedy zajistit

$$P\left(\left|\frac{1}{n}X - p\right| < 0,05\right) \simeq 0,9$$

tím, že zvolíme dostatečně veliký počet dotázaných studentů n.

Nyní můžeme přibližně počítat

$$\begin{aligned} 0,9 &\simeq P\left(\left|\frac{1}{n}X - p\right| < 0,05\right) = \\ &= P\left(-\frac{0,05n}{\sqrt{np(1-p)}} < \frac{X-np}{\sqrt{np(1-p)}} < \frac{0,05n}{\sqrt{np(1-p)}}\right) \simeq \\ &\simeq \Phi\left(\frac{0,05n}{\sqrt{np(1-p)}}\right) - \Phi\left(-\frac{0,05n}{\sqrt{np(1-p)}}\right) = \\ &= 2\Phi\left(\frac{0,05n}{\sqrt{np(1-p)}}\right) - 1. \end{aligned}$$

Podle Moivreovy-Laplaceovy věty lze pro velké n distribuční funkci aproximovat distribuční funkcí Φ normálního rozdělení. Proto

$$P\left(|Z| \le \frac{0{,}34}{3}\sqrt{n}\right) \approx \Phi\left(\frac{0{,}34}{3}\sqrt{n}\right) - \Phi\left(-\frac{0{,}34}{3}\sqrt{n}\right) = \\ = 2\Phi\left(\frac{0{,}34}{3}\sqrt{n}\right) - 1.$$

Celkem tedy dostáváme podmínku

$$2\Phi\left(\frac{0{,}34}{3}\sqrt{n}\right) - 1 \ge 0{,}995.$$

Odtud vypočítáme $n \ge \left(\frac{3z(0{,}9975)}{0{,}34}\right)^2 \approx 615$. □

9.72. Životnost (v hodinách) určité elektrické součástky má exponenciální rozdělení s parametrem $\lambda = \frac{1}{10}$. Pomocí centrální limitní věty odhadněte pravděpodobnost, že celková životnost 100 takových součástek bude mezi 900 a 1050 hodinami.

Řešení. V příkladu ‖9.51‖ jsem spočítali, že střední hodnota a rozptyl náhodné veličiny X_i s exponenciálním rozdělením jsou rovny $\mathrm{E}\,X_i = \frac{1}{\lambda}$ a $\mathrm{var}\,X_i = \frac{1}{\lambda^2}$. Střední životnost každé z našich součástek je tedy $\mathrm{E}\,X_i = \mu = 10$ hodin s rozptylem $\mathrm{var}\,X_i = \sigma^2 = 100$. Podle centrální limitní věty se rozdělení transformované náhodné veličiny $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right) = \frac{1}{100}\sum_{i=1}^{100} X_i - 10$ pro rostoucí n blíží normovanému normálnímu rozdělení. Proto hledanou pravděpodobnost pro životnost 100 součástek

$$P(900 \le \sum X_i \le 1050) = P\left(-1 \le \frac{1}{100}\sum_{i=1}^{100} X_i - 10 \le 0,5\right)$$

můžeme aproximovat pomocí distribuční funkce normálního rozdělení

$$P(900 \le \sum X_i \le 1050) \approx \Phi(0{,}5) - \Phi(-1) \approx 0{,}533.$$ □

9.73. Do bedny ukládáme výrobky se střední hodnotou 3 kg a směrodatnou odchylkou 0,8 kg. Jaký maximální počet výrobků můžeme do bedny uložit, aby celková hmotnost s pravděpodobností 99% nepřekročila jednu tunu?

Řešení. Označíme-li náhodnou veličinu, udávající hmotnost i-tého výrobku X_i, pak ze zadání $\mu = \mathrm{E}\,X_i = 3$ a $\sigma = \sqrt{\mathrm{var}\,X_i} = 0{,}8$ (vše v kg) a má platit

$$P(\sum_{i=1}^{n} X_i \le 1000) = 0{,}99.$$

Podle centrální limitní věty 9.42 lze rozdělení náhodné veličiny

$$S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{X_i - 3}{0{,}8}\right) = \frac{1}{0{,}8\sqrt{n}}\sum_{i=1}^{n} X_i - \frac{3\sqrt{n}}{0{,}8}$$

Chceme tedy dosáhnout

$$\Phi\left(\frac{0{,}05n}{\sqrt{np(1-p)}}\right) \simeq \frac{1}{2}(1 + 0{,}9) = 0{,}95.$$

Tento požadavek vede na volbu (připomeňme definici kritických hodnot $z(\alpha)$ pro veličinu s normovaným normálním rozdělením Z v odstavci 9.33)

$$\Phi\left(\frac{0{,}05n}{\sqrt{np(1-p)}}\right) \simeq z(0{,}05) = 1{,}64485.$$

Protože $p(1-p)$ nabývá největší hodnoty $\frac{1}{4}$, můžeme odtud odhadnout potřebný počet $n > 270$ nezávisle na p.

9.44. Přehled charakteristik některých rozdělení. V dalším se

vrátíme ke statistice a jistě nás nepřekvapí, že budeme pracovat s charakteristikami náhodných vektorů, které budou obdobné výběrovému průměru a rozptylu, ale také s relativními poměry takových charakteristik atd. Podíváme se proto teď na několik takových případů předem.

Uvažme náhodnou veličinu $Z \sim \mathrm{N}(0, 1)$ a spočtěme hustotu $f_Y(x)$ náhodné veličiny $Y = Z^2$. Evidentně je $f_Y(x) = 0$ pro $x \le 0$, zatímco pro kladná x

$$F_Y(x) = P(Y < x) = P(-\sqrt{x} < Z < \sqrt{x}) = \\ = \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-z^2/2}\,dz = \int_0^x \frac{1}{\sqrt{2\pi}}\,t^{-1/2}\,\mathrm{e}^{-t/2}\,dt.$$

Hustotu dostaneme derivací

$$f_Y(x) = \frac{d}{dx}F_Y(x) = \frac{1}{\sqrt{2\pi}}x^{-1/2}\,\mathrm{e}^{-x/2}.$$

Tomuto rozdělení se říká χ^2 *s jedním stupněm volnosti*, píšeme $Y \sim \chi^2$.

Budeme pracovat se součty takovýchto nezávislých veličin, ty ale všechny padnou do obecné třídy rozdělení s podobnými hustotami tvaru

$$f_X(x) = cx^{a-1}\,\mathrm{e}^{-bx}$$

pro $x > 0$, zatímco $f_X(x) = 0$ pro nekladná x, tj. naše rozdělení χ^2 odpovídá volbě $a = b = 1/2$. Tento případ jsme již podrobně diskutovali jako příklad v odstavci 9.25 a proto již víme, že taková funkce bude hustotou pro konstantu $c = \frac{b^a}{\Gamma(a)}$. Jde tedy o rozdělení $\Gamma(a, b)$ s hustotou pro kladná x

$$f_X(x) = \frac{b^a}{\Gamma(a)}x^{a-1}\,\mathrm{e}^{-bx}.$$

Obecně lze snadno spočíst k-tý moment takové veličiny X:

$$\mathrm{E}\,X^k = \int_0^\infty x^k\,\frac{b^a}{\Gamma(a)}x^{a-1}\,\mathrm{e}^{-bx}\,dx = \\ = \frac{\Gamma(a+r)}{\Gamma(a)b^r}\int_0^\infty x^k\,\frac{b^{a+r}}{\Gamma(a+r)}x^{a-1+r}\,\mathrm{e}^{-bx}\,dx = \\ = \frac{\Gamma(a+r)}{\Gamma(a)b^r},$$

protože integrál z hustoty rozdělení $\Gamma(a+r, b)$ v posledním upravovaném výrazu je nutně roven jedné.

Zejména tedy vidíme, že $\mathrm{E}\,X = \frac{\Gamma(a+1)}{b\Gamma(a)} = \frac{a}{b}$, zatímco

$$\mathrm{var}\,X = \frac{\Gamma(a+2)}{b^2\Gamma(a)} - \frac{a^2}{b^2} = \frac{(a+1)a - a^2}{b^2} = \frac{a}{b^2}.$$

aproximovat normovaným normálním rozdělením, a proto

$$P(\sum_{i=1}^{n} X_i \le 1000) = P(S_n \le \frac{1000}{0{,}8\sqrt{n}} - \frac{3\sqrt{n}}{0{,}8}) \approx \Phi(\frac{1000}{0{,}8\sqrt{n}} - \frac{3\sqrt{n}}{0{,}8}).$$

Z tabulek najdeme $z(0{,}99) \approx 2{,}326$, takže pro hledané n dostáváme kvadratickou rovnici

$$\frac{1000}{0{,}8\sqrt{n}} - \frac{3\sqrt{n}}{0{,}8} = 2{,}326,$$

ze které vypočítáme $n \approx 322$. □

I. Testování výběrů z normálního rozdělení

V 9.50 jsme se seznámili s tak zvaným oboustranným intervalovým odhadem neznámého parametru μ normálního rozložení $\mathrm{N}(\mu, \sigma^2)$. V některých případech nás zajímá pouze horní nebo dolní odhad, tj. statistika U respektive L, pro niž $P(\mu < U)$ respektive $P(L < \mu)$. Mluvíme pak o jednostranném intervalu spolehlivosti $(-\infty, U)$ respektive (L, ∞). Vztah pro výpočet těchto intervalů se odvodí obdobně jako u oboustranného intervalu. Pro náhodnou veličinu $Z = \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \sim \mathrm{N}(0, 1)$ tentokrát máme

$$1 - \alpha = \Phi(z(1-\alpha)) = P(Z < z(1-\alpha)).$$

Odtud okamžitě

$$1 - \alpha = P(\bar{X} - \frac{\sigma}{\sqrt{n}} z(1-\alpha) < \mu),$$

tedy $L = \bar{X} - \frac{\sigma}{\sqrt{n}} z(1-\alpha)$. Obdobně zjistíme $U = \bar{X} + \frac{\sigma}{\sqrt{n}} z(1-\alpha)$ a pro rozdělení s neznámým rozptylem $\mu \ge \bar{X} - \frac{S}{\sqrt{n}} t_{n-1}(1-\alpha)$ a $\mu \le \bar{X} + \frac{S}{\sqrt{n}} t_{n-1}(1-\alpha)$.

Pokud potřebujeme odhadnout rozptyl σ^2 náhodného rozložení, pak stejně jako u odvození odhadu střední hodnoty využijeme větu 9.49. Tentokrát ovšem využijeme její druhou část, podle které má náhodná veličina $\frac{n-1}{\sigma^2} S^2$ rozložení χ^2. Okamžitě je pak vidět, že platí

$$1 - \alpha = P\left(\chi^2_{n-1}(\alpha/2) \le \frac{n-1}{\sigma^2} S^2 \le \chi^2_{n-1}(1-\alpha/2)\right).$$

Oboustranný $100(1-\alpha)\%$ interval spolehlivosti pro rozptyl je tedy

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1}(1-\alpha/2)}, \frac{(n-1)S^2}{\chi^2_{n-1}(\alpha/2)}\right)$$

a podobně pro jednostranný horní a dolní odhad dostaneme

$$\sigma^2 \le \frac{(n-1)S^2}{\chi^2_{n-1}(\alpha)}, \text{ resp. } \frac{(n-1)S^2}{\chi^2_{n-1}(1-\alpha)} \le \sigma^2.$$

Úplně obdobně spočteme momentovou vytvořující funkci pro všechny hodnoty $-b < t < b$

$$\begin{aligned} M_X(t) &= \int_0^\infty \mathrm{e}^{tx} \frac{b^a}{\Gamma(a)} x^{a-1} \mathrm{e}^{-bx}\, dx = \\ &= \frac{b^a}{(b-t)^a} \int_0^\infty x^k \frac{(b-t)^a}{\Gamma(a)} x^{a-1} \mathrm{e}^{-(b-t)x}\, dx = \\ &= \frac{b^a}{(b-t)^a}. \end{aligned}$$

Pro součet nezávislých rozdělení $Y = X_1 + \cdots + X_n$ s rozděleními $X_i \sim \Gamma(a_i, b)$ tedy okamžitě dostáváme momentovou vytvořující funkci (pro hodnoty $|t| < b$)

$$M_Y(t) = \left(\frac{b}{b-t}\right)^{a_1+\cdots+a_n},$$

tj. $Y \sim \Gamma(a_1 + \cdots + a_n, b)$. Velmi podstatný je ovšem přitom předpoklad, že všechna gamma rozdělení sdílí stejnou hodnotu b.

Jako okamžitý důsledek nyní dostáváme hustotu rozdělení veličiny $Y = Z_1^2 + \cdots + Z_n^2$, kde všechna $Z_i \sim \mathrm{N}(0, 1)$. Jde totiž podle právě dokázaného o gamma rozdělení $Y \sim \Gamma(n/2, 1/2)$ a má proto hustotu

$$f_Y(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} \mathrm{e}^{-x/2}.$$

Tomuto speciálnímu případu gamma rozdělení říkáme rozdělení χ^2 *s n stupni volnosti*. Značíme jej zpravidla $Y \sim \chi^2_n$.

9.45. F-rozdělení a t-rozdělení. Ve statistických úvahách často chceme porovnávat dva různé výběrové rozptyly a bude tedy třeba uvažovat veličiny, které jsou dány podílem

$$U = \frac{X/k}{Y/m},$$

přičemž $X \sim \chi^2_k$ a $Y \sim \chi^2_m$.

Budeme chtít spočíst hustotu takového rozdělení a začneme obecnější úvahou. Předpokládejme, že $f_X(x)$ a $f_Y(x)$ jsou hustoty nezávislých náhodných veličin X a Y a f_Y je nenulové pouze pro kladná x. Spočteme si distribuční funkci náhodné veličiny $U = cX/Y$, kde $c > 0$ je libovolná konstanta. Při výpočtu použijeme Fubiniho větu o záměnnosti integrování podle jednotlivých proměnných.

$$\begin{aligned} F_U(u) = P(X < (u/c)Y) &= \int_0^\infty \int_{-\infty}^{uy/c} f_X(x) f_Y(y)\, dx\, dy = \\ &= \int_0^\infty \left(\int_{-\infty}^{u} \frac{y}{c} f_X(ty/c) f_Y(y)\, dt\right) dy = \\ &= \int_{-\infty}^{u} \left(\frac{1}{c} \int_0^\infty y f_X(ty/c) f_Y(y)\, dy\right) dt. \end{aligned}$$

Z tohoto výrazu pro $F_U(u)$ okamžitě plyne, že hustota f_U náhodné proměnné U je rovna

$$f_U(u) = \frac{1}{c} \int_0^\infty y f_X(uy/c) f_Y(y)\, dy.$$

Když teď dosadíme hustoty příslušných speciálních gamma rozdělení za $X \sim \chi^2_k$ a $Y \sim \chi^2_m$ a za konstantu c zvolíme

9.74. Při 600 hodech kostkou padla šestka celkem 45 krát. Je možné tvrdit, že jde o ideální kostku na hladině $\alpha = 0{,}01$?

Řešení. Pro ideální kostku je pravděpodobnost hodu šestky při každém hodu rovna $p = \frac{1}{6}$. Počet šestek v 600 hodech je pak dán náhodnou veličinou X, která má binomické rozdělení $X \sim \mathrm{Bi}(600, \frac{1}{6})$. Toto rozdělení můžeme podle 9.42 aproximovat rozdělením $\mathrm{N}(100, \frac{250}{3})$. Naměřenou hodnotu $X = 45$ můžeme považovat za náhodný výběr o jednom členu. Pokládáme-li rozptyl za známý, pak podle 9.50 je pak 99% (oboustranný) interval spolehlivosti pro střední hodnotu μ roven $(45 - \sqrt{\frac{250}{3}}z(0{,}995),\ 45 + \sqrt{\frac{250}{3}}z(0{,}995))$. Z tabulek zjistíme, že kvantil přibližně $z(0{,}995) \approx 2{,}58$, což dává interval $(21, 69)$. Na ideální kostce je ale zřejmě $\mu = 100$, a proto nejde v tomto smyslu o ideální kostku na hladině $\alpha = 0{,}01$. □

9.75. Předpokládejme, že výška desetiletých chlapců má normální rozdělení $\mathrm{N}(\mu, \sigma^2)$ s neznámou střední hodnotou μ a rozptylem $\sigma^2 = 39{,}112$. Změřením výšky 15 chlapců jsme určili výběrový průměr $\bar{X} = 139{,}13$. Určete

i) 99% oboustranný interval spolehlivosti pro parametr μ,

ii) dolní odhad μ na hladině významnosti 95%.

Řešení. a) Podle 9.50 je $100(1-\alpha)\%$ oboustranný interval spolehlivosti pro neznámou střední hodnotu μ normálního rozložení dán výrazem

$$\mu \in \left(\bar{X} - \frac{\sigma}{\sqrt{n}}z(1-\alpha/2),\ \bar{X} + \frac{\sigma}{\sqrt{n}}z(1-\alpha/2)\right), \tag{9.3}$$

kde $\bar{X}$ je výběrový průměr z n hodnot, σ^2 je známý rozptyl a $z(1-\alpha/2)$ je příslušný kvantil. Přímým dosazením ze zadání $n = 15$, $\sigma \approx 6{,}254$ a z tabulek $z(0{,}995) \approx 2{,}576$ dostaneme $\frac{\sigma}{\sqrt{n}}z(\alpha/2) \approx 4{,}16$, tj. $\mu \in (134{,}97,\ 143{,}29)$.

b) Dolní odhad L parametru μ na hladině významnosti 95% je určen výrazem $L = \bar{X} - \frac{\sigma}{\sqrt{n}}z(0{,}95)$. Z tabulek $z(0{,}95) \approx 1{,}645$, a proto přímým dosazením dostáváme $\mu \in (136{,}474, \infty)$. □

9.76. Odběratel provádí kontrolu jakosti námi dodaných výrobků namátkovou kontrolou testovaného rozměru u 21 náhodně vybraných výrobků. Dodávka bude přijata, pokud nebude výběrová směrodatná odchylka překračovat hodnotu 0,2 mm. Víme přitom, že naše stroje produkují výrobky, u nichž má sledovaný rozměr normální rozdělení tvaru $\mathrm{N}(10\,\mathrm{mm};\ 0{,}0734\,\mathrm{mm}^2)$. S využitím statistických tabulek určete pravděpodobnost, s níž bude dodávka přijata. Jak se změní odpověď, pokud odběratel kvůli nákladům na testy začne testovat pouze 4 výrobky?

$c = m/k$, dostaneme pro náhodnou veličinu $U = \frac{X/k}{Y/m}$ hustotu $f_U(u)$

$$\frac{(k/m)^{k/2}}{2^{(k+m)/2}u^{k/2-1}\Gamma(k/2)\Gamma(m/2)}\int_0^\infty y^{(k+m)/2-1}\,\mathrm{e}^{-y(1+ku/m)/2}\,dy.$$

(Ověřte si sami!) Poslední integrál obsahuje, až na konstantní násobek hustotu rozdělení $\Gamma((k+m)/2, (1+ku/m)/2)$, takže hledaná hustota bude mít tvar

$$f_U(u) = \frac{\Gamma((k+m)/2)}{\Gamma(k/2)\Gamma(m/2)}\Big(\frac{k}{m}\Big)^{k/2}u^{k/2-1}\Big(1+\frac{k}{m}u\Big)^{-(k+m)/2}.$$

Takovému rozdělení se říká *Fisherovo-Snedecorovo rozdělení s k a m stupni volnosti*, zkráceně také *F-rozdělení*.

Další potřebné rozdělení se objevuje při zkoumání podílu veličin $Z \sim \mathrm{N}(0,1)$ a $\sqrt{X/n}$, kde $X \sim \chi^2_n$ (tj. zajímá nás poměr Z a směrodatné odchylky nějakého výběru).

Spočteme nejdříve opět snadno distribuční funkci pro $Y = \sqrt{X}$ (všimněme si, že X a tedy i Y nabývají s nenulovou pravděpodobností pouze kladných hodnot)

$$\begin{aligned} F_Y(y) &= P(\sqrt{X} < y) = P(X < y^2) = \\ &= \int_0^{y^2} \frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}\,\mathrm{e}^{-x/2}\,dy = \\ &= \int_0^{y} \frac{1}{2^{n/2-1}\Gamma(n/2)}t^{n-1}\,\mathrm{e}^{-t^2/2}\,dt. \end{aligned}$$

Odtud již vidíme, že hustota náhodné veličiny Y je

$$f_Y(y) = \frac{1}{2^{n/2-1}\Gamma(n/2)}y^{n-1}\,\mathrm{e}^{-y^2/2}.$$

Nyní můžeme použít stejný postup jako v předchozím odstavci u náhodné veličiny $U = cZ/Y$ a volíme $c = \sqrt{n}$, $Y = \sqrt{X}$. Dostaneme tedy pro náhodnou veličinu

$$T = \frac{Z}{\sqrt{X/n}}$$

po krátkém výpočtu, podobném jako výše, hustotu $f_T(t)$ ve tvaru

$$f_T(t) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{n\pi}}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}.$$

Tomuto rozdělení říkáme *Studentovo t-rozdělení s n stupni volnosti*.

9.46. Vícerozměrné normální rozdělení. Jestliže má náhodný vektor $Z = (Z_1, \dots, Z_n)$ nezávislé komponenty $Z_i \sim \mathrm{N}(0,1)$, je jeho varianční matice jednotkovou maticí, tj. $\mathrm{var}\,Z = \mathbb{I}_n$.

Často ale potkáváme v praktických problémech náhodné vektory, které z takového vektoru Z vznikají obecnou afinní transformací $U = a + BZ$, kde a je libovolný konstantní vektor v $\mathbb{R}^m$ a B je konstantní matice typu (m, n).

Jak jsme odvodili ve větách 9.32 a 9.38, takové náhodné vektory mají střední hodnotu $\mathrm{E}\,U = a$ a varianční matici $\mathrm{var}\,U = V = BB^T$ (protože varianční matice Z je identická). Je tedy tato varianční matice vždy pozitivně semidefinitní.

Říkáme, že náhodný vektor U má *mnohoměrné normální rozdělení* $\mathrm{N}_m(a, V)$.

Pro libovolné mnohoměrné normální rozdělení $\mathrm{N}_m(a, V)$ můžeme znovu uvážit afinní transformaci

$$W = c + DU$$

Řešení. Podle zadání hledáme pravděpodobnost $P(S \leq 0{,}2)$. Využijeme větu 9.49, podle které má při náhodném výběru n výrobků náhodná veličina $\frac{n-1}{\sigma^2}S^2$ rozdělení χ^2_{n-1}. V našem případě $n = 21$ a $\sigma^2 = 0{,}0734$, a proto

$$P(S \leq 0{,}2) = P\left(\frac{20}{0{,}0734}S^2 \leq \frac{20}{0{,}0734}0{,}2^2\right) = \chi^2_{20}\left(\frac{20 \cdot 0{,}2^2}{0{,}0734}\right)$$

Výraz v argumentu distribuční funkce je roven přibližně 10,9 a z tabulek pro χ^2 rozložení zjistíme $\chi^2_{20}(10{,}9) \approx 0{,}05$. Pravděpodobnost , že odběratel dodávku přijme je tedy pouze 5%. To, že tato pravděpodobnost bude malá lze odvodit i bez počítání, platí totiž $\mathrm{E}\,S^2 = \sigma^2 = 0{,}0734 > 0{,}2^2$. Pokud bude odběratel testovat pouze 4 výrobky, pak je zřejmě pravděpodobnost přijetí dodávky dána výrazem $\chi^2_3\left(\frac{3 \cdot 0{,}2^2}{0{,}0734}\right) \approx \chi^2_3(1{,}63)$. Hodnotu distribuční funkce χ^2 v tomto argumentu nelze ve většině statistických tabulek nalézt. Proto ji odhadneme lineární interpolací. Jsou-li například nejbližší body $\chi^2_3(0{,}58) = 0{,}1$ a $\chi^2_3(6{,}25) = 0{,}9$, pak

$$\chi^2_3(1{,}63) \approx (1{,}63 - 0{,}58)\frac{0{,}9 - 0{,}1}{6{,}25 - 0{,}58} + 0{,}1 \approx 0{,}24.$$

Tento výsledek je sice jen odhad, ale určitě bude pravděpodobnost přijetí dodávky v případě testování 4 výrobků výrazně vyšší než v předchozím případě. □

9.77. Ze základního souboru, z rozdělení $N(\mu, \sigma^2)$, kde $\sigma^2 = 0{,}06$ jsme pořídili náhodný výběr s realizacemi 1,3; 1,8; 1,4; 1,2; 0,9; 1,5; 1,7. Určete oboustranný 95% interval spolehlivosti pro neznámou střední hodnotu.

Řešení. Ze zadání se jedná o náhodný výběr rozsahu $n = 7$ z normálního rozložení se známým rozptylem $\sigma^2 = 0{,}06$. Výběrový průměr je

$$\bar{X} = \frac{1}{7}(1{,}3 + 1{,}8 + 1{,}4 + 1{,}2 + 0{,}9 + 1{,}5 + 1{,}7) = 1{,}4$$

a z tabulek pro danou hladinu spolehlivosti $\alpha = 0{,}05$ zjistíme $z(1 - \alpha/2) = z(0{,}975) \approx 1{,}96$. Dosazením do (‖9.3‖) pak ihned dostaneme hledaný interval (1,22, 1,58). □

9.78. Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení $N(\mu, 0{,}04)$. Určete nejmenší počet měření, který je třeba provést, aby šířka 95% intervalu spolehlivosti pro μ nepřesáhla 0,16.

Řešení. Protože se jedná o normální rozložení se známým rozptylem, je šířka $(1 - \alpha)$% intervalu spolehlivosti je podle (‖9.3‖) rovna $\frac{2\sigma}{\sqrt{n}}z(1 - \alpha/2)$. Dosazením hodnot ze zadání tedy dostaneme pro počet měření n nerovnici

$$\frac{2 \cdot 0{,}2}{\sqrt{n}}z(0{,}975) \leq 0{,}16.$$

s vektorem konstant $c \in \mathbb{R}^k$ a libovolnou konstantní maticí typu (k, m). Přímým výpočtem vidíme, že

$$W = c + D(a + BZ) = (c + Da) + (DB)Z,$$

což je samozřejmě náhodný vektor $W \sim \mathrm{N}_k(c + Da, DB^T BD^T)$. Chová se tedy kovarianční matice mnohoměrného normálního rozdělení při afinních transformacích jako kvadratická forma.

Tato přímočará úvaha ukazuje, že jakákoliv lineární kombinace komponent složek náhodného vektoru s mnohoměrným normálním rozdělením je náhodná veličina s normálním rozdělením. Stejně je každý vektor vzniklý výběrem jen některých komponent vektoru U opět náhodným vektorem s mnohoměrným normálním rozdělením.

Poznamenejme závěrem, že když pro transformaci náhodného vektoru $Z \sim \mathrm{N}_n(0, \mathbb{I}_n)$ použijeme ortogonální transformaci s maticí Q^T, pak můžeme přímo spočíst sdruženou distribuční funkci náhodného vektoru $U = Q^T Z$. Skutečně, jestliže transformaci budeme v souřadnicích psát jako $t = Q^T z$, pak její inverze je $z = Qt$ a Jakobián této transformace je roven jedné. Proto (všimněme si že také jistě platí $\sum_i z_i^2 = \sum_i t_i^2$)

$$\begin{aligned}
F_U(u) &= P(U_i < u_i, i = 1, \ldots, n) = \\
&= \int \cdots \int_{z;\, Q^T z < u} (2\pi)^{-n/2}\, \mathrm{e}^{-\sum z_i^2/2}\, dz_1 \cdots dz_n = \\
&= \int \cdots \int_{t;\, t < u} (2\pi)^{-n/2}\, \mathrm{e}^{-\sum t_i^2/2}\, dt_1 \cdots dt_n = \\
&= \left(\int_{-\infty}^{u_1} (2\pi)^{-1/2}\, \mathrm{e}^{-t_1^2}\, dt_1\right) \cdots \\
&\quad \cdots \left(\int_{-\infty}^{u_n} (2\pi)^{-1/2}\, \mathrm{e}^{-t_n^2}\, dt_1\right) = \\
&= F_{U_1}(u_1) \cdots F_{U_n}(u_n)
\end{aligned}$$

Odtud okamžitě plyne, že všechny komponenty náhodného vektoru U jsou opět nezávislé a opět je $U \sim \mathrm{N}_n(0, \mathbb{I}_n)$.

3. Matematická statistika

Jakkoli je zpracování dat v matematické statistice založené na velmi sofistikované matematice, skutečné aplikace již matematiku jako vědu dalece přesahují a vždy jsou založeny také na vstupech z těch oborů, pro které má být použití podstatné.

I proto se omezíme v této učebnici jen na skromné poznámky o statistických metodách a postupech a odkazujeme zájemce na volbu speciální literatury (odrážející i zamýšlené oblasti aplikace).

9.47. Přípravné úvahy. V popisné statistice jsme se na začátku kapitoly snažili datové soubory opatřit charakteristikami, které nám o nich vypovídaly podstatné údaje typu výběrového průměru, rozptylu apod.

Matematická statistika pracuje s nějakým výběrem z daného základního souboru a snaží se postihnout, do jaké míry jsou zjištěné statistiky relevantní, případně se ze zjištěných dat pokouší zjistit nebo upřesnit vhodný teoretický model pro chování celého souboru (a z něj pak třeba odhadovat pravděpodobnost nějakého budoucího jevu).

Protože $z(0{,}975) \approx 1{,}96$, dostáváme odtud $n \geq 24{,}01$. Je tedy třeba provést aspoň 25 pokusů. □

9.79. Náhodná veličina X má normální rozdělení $N(\mu, \sigma^2)$, kde μ, σ^2 nejsou známy. V následující tabulce jsou uvedeny četnosti jednotlivých realizací této náhodné veličiny.

X_i	8	11	12	14	15	16	17	18	20	21
n_i	1	2	3	4	7	5	4	3	2	1

Vypočtěte výběrový průměr, výběrový rozptyl, výběrovou směrodatnou odchylku a určete 99% interval spolehlivosti pro střední hodnotu μ.

Řešení. Výběrový průměr je dán výrazem $\bar{X} = \sum n_i X_i / \sum n_i$. Dosazením hodnot ze zadání máme $\bar{X} = 490/32 \approx 15{,}3$. Výběrový rozptyl je z definice $S = \sum n_i (X_i - \bar{X})^2 / (\sum n_i - 1)$. Po dosazení daných hodnot dostaneme $S^2 = 1943/256 \approx 7{,}6$, a proto výběrová směrodatná odchylka splňuje $S \approx 2{,}8$. Vzorec pro oboustranný $(1-\alpha)$% interval spolehlivosti pro střední hodnotu μ při neznámém rozptylu jsme odvodili na konci části 9.50

$$\mu \in \left(\bar{X} - \frac{S}{\sqrt{n}} t_{n-1}(1-\alpha/2),\, \bar{X} + \frac{S}{\sqrt{n}} t_{n-1}(1-\alpha/2) \right).$$

Přímým dosazením $\bar{X} = 15{,}3$, $n = 32$, $S \approx 2{,}8$, $\alpha = 0{,}01$ a z tabulek $t_{31}(0{,}995) \approx 2{,}75$ pak zjistíme, že 99% interval spolehlivosti je $\mu \in (14{,}0, 16{,}7)$. □

9.80. Pomocí přiložené tabulky distribuční funkce standardního normálního rozdělení určete pravděpodobnost, že při 3600 hodech mincí bude rozdíl mezi počtem padlých hlav a orlů větší než 90.

Uvažme jednoduchý příklad, kdy si sami zhotovíme dřevěnou minci s rubem a lícem. Hodíme jí n-krát a víme, že přitom padlo $k \leq n$ líců. Chceme z tohoto experimentu vyvodit závěr, s jakou pravděpodobností v dalších dvou hodech padne vždy líc.

K této úloze můžeme mít dva základní přístupy. Jedním je tzv. klasická statistika (neboli *frekvenční statistika*). Vyjdeme z předpokladu, že jednotlivé hody jsou nezávislé a ve všech je stejná pravděpodobnost líce dána objektivně existujícím parametrem $\theta = p$ (který jen dosud neznáme). Jednotlivé hody tedy považujeme za realizaci náhodné veličiny X s alternativním rozdělením pravděpodobnosti. Pravděpodobnost, že padlo k líců z n pokusů je dána binomiálním rozdělením a lze očekávat že „nejlepší možný“ odhad parametru p bude dán poměrem $\theta = k/n$. Obvyklým cílem je pak opatřit takový odhad vyjádřením o jeho spolehlivosti, který můžeme odvinout od znalosti celkového počtu pokusů n a znalosti asymptotického chování modelu při rostoucím n. Jestliže tedy např. padne 8 líců z 10 pokusů, budeme s jistou (matematicky odhadnutou) spolehlivostí tvrdit, že pravděpodobnost dvou následujících líců bude $0{,}8^2 = 0{,}64$, tj. výrazně více než polovina.

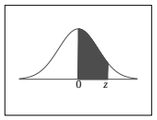
Druhou možností je postupovat víceméně naopak. Můžeme totiž považovat parametr θ za náhodnou proměnnou, data získaná experimentem za konstanty a pokoušet se z nich vydedukovat informace o rozložení pravděpodobnosti této náhodné veličiny θ. Vycházíme přitom z nějakých vstupních informací o tomto rozložení. Jestliže tedy např. budeme předpokládat, že mince vznikla z homogenního materiálu vcelku přesným soustružením a následným rozlišením lícu a rubu barevným nátěrem, můžeme jako vstupní předpoklad o θ použít rovnoměrné rozdělení pravděpodobnosti rozložené na malinkém intervalu odpovídajím přesnosti soustruhu. Pak ovšem lze očekávat, že stejný experiment také povede k vychýlení odhadu pravděpodobnosti dvou následujících líců od hodnoty $0{,}5^2 = 0{,}25$ pro dokonalou minci, půjde ale patrně o poněkud menší pravděpodobnost než v předchozím postupu. Hovoříme tu o tzv. *bayesovské statistice*.

První přístup vychází z ryze matematické abstrakce, že pravděpodobnosti jsou dány četnostmi výskytů jevů v tak velkých vzorcích dat, že je můžeme dobře aproximovat nekonečnými modely a využít pro odhady spolehlivosti centrální limitní věty. Statistik zde na pravděpodobnost pohlíží jako na idealizaci relativní četnosti případů, v nichž se vyskytne určitý výsledek při opakovaných pokusech. Tato zdánlivá výhoda/rigorózност se může ale rychle stát nevýhodou, jakmile se začneme zabývat spolehlivostí samotných dat a vhodností zvoleného experimentu. Stejně tak je obtížné frekvenční statistiku dobře použít pro odhad pravděpodobnosti výskytu jednorázového děje.

Bayesovská statistika je naopak příkladem matematizace „selského rozumu“, když chceme naše původní přesvědčení postupně pozměňovat ve světle nových dat.

Je zajímavé, že historicky byl zjevně první bayesovský přístup (např. Laplace a další již v 18. století), který byl prakticky zcela vystřídán frekvenční statistikou ve 20. století. V posledních desetiletích se však ale bayesovská statistika vrátila, společně s dalšími novými přístupy, do popředí zájmu.

Standard Normal Distribution Table



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998

Gilles Cazelais. Typeset with LaTeX on April 20, 2006.

Řešení. Označíme-li jako X náhodnou veličinu udávající počet padlých hlav, tak X má binomické rozložení pravděpodobnosti $Bi(3600, 1/2)$ (se střední hodnotou 1800 a směrodatnou odchylkou 30) a tudíž lze distribuční funkci veličiny $\frac{X-1800}{30}$ lze pro dané velké $n = 3600$ podle Moivreovy-Laplaceovy věty velmi dobře odhadnout jako distribuční funkci Φ standardního normálního rozdělení. Hledaná pravděpodobnost je tedy

$$P = 1 - P[1755 \le X \le 1845] =$$
$$= 1 - P\left[-1,5 \le \frac{X-1800}{30} \le 1,5\right] = 2\Phi(-1{,}5) \doteq 0{,}1336,$$

kde poslední hodnotu jsme zjistili z přiložené tabulky. □

9.81. Pravděpodobnost narození chlapce je 0,515. Jaká je pravděpodobnost, že mezi deseti tisíci novorozenci bude stejně nebo více děvčat než chlapců.

Řešení.

$$P[X < 5000] = P[\underbrace{\frac{X-5150}{\sqrt{5150 \cdot 0{,}485}}}_{\sim N(0,1)} < \underbrace{\frac{-150}{\sqrt{5150 \cdot 0{,}485}}}_{-3{,}001\ldots}] \doteq$$
$$\doteq 0{,}00135$$

□

9.48. Náhodný výběr z populace. Budeme se nejprve zabývat

prvním přístupem z předchozího odstavce. Předpokládejme tedy, že máme k dispozici (velký) základní statistický soubor s N jednotkami, který nazýváme *populace*, a zároveň nějaký číselný znak pro každou z jednotek, tj. soubor hodnot $(x_1, \ldots, x_N)$. Z něj ovšem máme k dispozici pouze *výběrový soubor* s hodnotami $(X_1, \ldots, X_n)$.

Abychom se vyhnuli diskusi skutečné velikosti základního statistického souboru s N jednotkami, budeme předpokládat, že vybíráme položky výběrového souboru jednu po druhé a každou vybranou jednotku poté do populace vracíme. Zároveň předpokládáme, že každá položka má stejnou pravděpodobnost výběru $1/N$. Hovoříme pak o *náhodném výběru*.

Způsob realizace náhodného výběru nyní interpretujeme tak, že pracujeme s vektorem $(X_1, \ldots, X_n)$ nezávislých náhodných veličin a že všechny tyto veličiny mají stejné rozdělení pravděpodobnosti. Zejména tedy budou sdílet distribuční funkci $F_X(x)$ a momenty

$$\mathrm{E}\, X_i = \mu, \quad \mathrm{var}\, X_i = \sigma^2.$$

Dalším naším krokem musí být odvození charakteristik výběrového průměru $\bar{X}$ a *výběrového rozptylu*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

přičemž následující věta dává hned zdůvodnění, proč volíme koeficient $\frac{1}{n-1}$ místo $\frac{1}{n}$, jak tomu bylo u s^2 v odstavci 9.6.

Věta. *Pro výběrový průměr $\bar{X}$ spočítaný z náhodného výběru rozsahu n z rozdělení s konečnou střední hodnotou μ a konečným rozptylem σ^2 platí*

$$\mathrm{E}\,\bar{X} = \mu, \quad \mathrm{var}\,\bar{X} = \frac{1}{n}\sigma^2.$$

Pro výběrový rozptyl S^2 platí

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

Důkaz. Jak jsme odvodili v odstavci 9.32, je

$$\mathrm{E}\,\bar{X} = \frac{1}{n}\mathrm{E}\sum_{i=1}^{n} X_i = \frac{1}{n}n\mu = \mu.$$

Díky nezávislosti veličin X_i můžeme použít aditivnost rozptylu odvozenou v odstavci 9.36 a viděli jsme také, že vůči násobení skalárem se rozptyl chová jako kvadratická forma. Dostáváme proto

$$\mathrm{var}\,\bar{X} = \frac{1}{n^2}\,\mathrm{var}\sum_{i=1}^{n} X_i = \frac{1}{n^2}n\sigma^2 = \frac{1}{n}\sigma^2.$$

Přímým roznásobením se ověří vztah

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2.$$

9.82. Pomocí distribuční funkce standardního normálního rozdělení určete pravděpodobnost, že při 18000 hodech šestibokou kostkou padne alespoň 3100 šestek.

Řešení. Obdobně, jako v předchozích příkladech. X má binomické rozdělení pravděpodobnosti $Bi(18000, 1/6)$. Určíme střední hodnotu $((1/6)(18000) = 3000)$, směrodatnou odchylku $\sqrt{((1/6)(1-1/6)18000)} = 50$, tedy veličinu $\frac{X-3000}{50}$ lze odhadnout jako distribuční funkci Φ standardního normálního rozložení:

$$P[X \geq 3100] = P\left[\frac{X-3000}{50} \geq \frac{3100-3000}{50}\right] = \\ = P\left[\frac{X-3000}{50} \geq 2\right] \doteq 1 - \Phi(2) \doteq 0{,}0228.$$

□

9.83. Agentura pro výzkum veřejného mínění pořádá průzkum volebních preferencí pěti vybraných politických stran. Kolik náhodně vybraných respondentů se musí výzkumu zúčastnit, aby byly s pravděpodobností 0,95 výsledky průzkumu byly u všech zkoumaných stran v rozmezí ±2% od skutečných preferencí?

Řešení. Nechť $p_i, i = 1 \ldots 5$ je skutečná relativní četnost příznivců i-té politické strany v populaci a nechť náhodná veličina X_i udává počet příznivců této strany mezi náhodně zvolenými n voliči. Budeme považovat za nezávislé jevy, že do daného intervalu padne X_i/n. Pokud zvolíme n takové, že pro všechna i padne X_i/n do daného intervalu s pravděpodobností alespoň $\sqrt[5]{0{,}95} \doteq 0{,}99$, bude požadavek zadání splněn. Hledejme tedy n takové, že $P[|\frac{X}{n} - p| < 0{,}02] \geq 0{,}99$. Nejprve upravme vyjádření hledané pravděpodobnosti:

$$\begin{aligned} & P\left[\left|\frac{X}{n} - p\right| < 0{,}02\right] \\ &= P\left[-0{,}02 < \frac{X}{n} - p < 0{,}02\right] = \\ &= P\left[-0{,}02 \cdot n < X - pn < 0{,}02 \cdot n\right] = \\ &= P\left[\frac{-0{,}02 \cdot n}{\sqrt{np(1-p)}} < \frac{X-pn}{\sqrt{np(1-p)}} < \frac{0{,}02 \cdot n}{\sqrt{np(1-p)}}\right] = \\ &= \Phi\left(\frac{0{,}02 \cdot n}{\sqrt{np(1-p)}}\right) - \Phi\left(-\frac{0{,}02 \cdot n}{\sqrt{np(1-p)}}\right) = \\ &= 2\Phi\left(\frac{0{,}02 \cdot n}{\sqrt{np(1-p)}}\right) - 1, \end{aligned}$$

Můžeme tedy spočíst:

$$\mathrm{E}\, s^2 = \frac{1}{n}\mathrm{E}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{n}{n}(\bar{X} - \mu)^2 = \\ = \frac{1}{n}\sum_{i=1}^{n}\operatorname{var} X_j - \operatorname{var}\bar{X} = \\ = \left(1 - \frac{1}{n}\right)\sigma^2.$$

Proto upravujeme rozptyl s^2 vynásobením koeficientem $\frac{n}{n-1}$ a dostáváme právě výběrový rozptyl S^2 a jeho střední hodnotu σ. Tato poslední úprava samozřejmě nemá smysl pro $n = 1$. □

9.49. Náhodný výběr z normálního rozdělení. V praktických úlohách je třeba znát nejen číselné charakteristiky výběrového průměru a rozptylu, ale jejich úplné rozdělení pravděpodobnosti. To můžeme samozřejmě odvodit, pouze známe-li konkrétní rozdělení pravděpodobnosti X_i. Jako užitečnou ilustraci si spočtěme výsledek pro náhodný výběr z normálního rozdělení.

Již jsme ověřili jako příklad na vlastnosti momentových vytvořujících funkcí v 9.40, že součet náhodných veličin s normálními rozděleními je opět normální rozdělení. Odtud je zřejmé, že i výběrový průměr musí mít normální rozdělení a protože již známe jeho střední hodnotu a rozptyl, bude $\bar{X} \sim \mathrm{N}(\mu, \frac{1}{n}\sigma^2)$.

O něco složitější je to s odvozením rozdělení pravděpodobnosti výběrového rozptylu. Tady si pomůžeme úvahami o mnohoměrných normálních rozděleních z odstavce 9.40. Uvažme vektor Z normovaných normálních veličin

$$Z_i = \frac{X_i - \mu}{\sigma}.$$

Stejnou vlastnost má i vektor $U = Q^T Z$ s jakoukoliv ortogonální maticí Q. Vždy přitom také platí $\sum_i U_i^2 = \sum_i X_i^2$. Zvolíme si takovou matici Q, aby první komponenta U_1 byla, až na násobek, rovna výběrovému průměru $\bar{Z}$. Tzn. zvolíme si první sloupec matice Q ve tvaru $(\sqrt{n})^{-1}(1, \ldots, 1)$. Pak tedy $U_1^2 = n\bar{Z}^2$ a můžeme počítat:

$$\sum_{i=1}^{n} U_i^2 = \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n}(Z_i - \bar{Z})^2 + n\bar{Z}^2$$

$$\sum_{i=2}^{n} U_i^2 = \sum_{i=1}^{n}(Z_i - \bar{Z})^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Je tedy násobek výběrového rozptylu $\frac{n-1}{\sigma^2}S^2$ součtem $n-1$ kvadrátů normalizovaných normálních veličin a dokázali jsme následující tvrzení:

Věta. *Je-li $(X_1, \ldots, X_n)$ náhodný výběr z rozdělení $\mathrm{N}(\mu, \sigma^2)$, pak jsou $\bar{X}$ a S^2 nezávislé veličiny a platí*

$$\bar{X} \sim \mathrm{N}(\mu, \frac{1}{n}\sigma^2), \quad \frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}.$$

Okamžitým důsledkem je, že normalizovaný výběrový průměr

$$T = \sqrt{n}\frac{\bar{X} - \mu}{S}$$

má studentovo t-rozdělení pravděpodobnosti s $n-1$ stupni volnosti.

kde Φ je distribuční funkce normálního rozdělení. Řešme tedy nerovnici:

$$\begin{aligned} 2\Phi\left(\frac{0{,}02\cdot n}{\sqrt{np(1-p)}}\right)-1 &\geq 0{,}99 \\ \Phi\left(\frac{0{,}02\cdot n}{\sqrt{np(1-p)}}\right) &\geq 0{,}995 \end{aligned}$$

Protože distribuční funkce je rostoucí je poslední podmínka ekvivalentní

$$\begin{aligned} \frac{0{,}02\cdot n}{\sqrt{np(1-p)}} &\geq \Phi^{-1}(0{,}995) \\ \frac{0{,}02\cdot n}{\sqrt{np(1-p)}} &\geq 2{,}576 \\ \sqrt{n} &\geq 50\cdot 2{,}576\cdot \underbrace{\sqrt{p(1-p)}}_{\leq \frac{1}{2}} \Longrightarrow \\ \Longrightarrow n &\geq (25\cdot 2{,}276)^2\cdot 4147 \end{aligned}$$

Při tom jsme použili faktu, že funkce $p(1-p)$ nabývá svého maxima pro $p=\frac{1}{2}$ a tímto maximem je $\frac{1}{4}$. Vidíme, že pokud např. $p \doteq 0{,}1$, pak je $\sqrt{p(1-p)}=0{,}3$ a hodnota minimálního n je menší. To odpovídá očekávání: k odhadu méně populárních stran, stačí méně respondentů (pokud agentura odhadne zisk takové strany jako 2% bez toho, aniž by se někoho ptala, tak má požadovanou přesnost téměř jistě zaručenu).

□

9.84. Dvouvýběrový test. Uvažme dva náhodné vektory Y_1 a Y_2, jejichž všechny složky jsou po dvou nezávislé náhodné veličiny s normálním rozdělením, a předpokládejme, že složky vektoru Y_i mají stejnou střední hodnotu μ_i, zatímco rozptyl σ je stejný pro všechny komponenty.

Použijte obecný lineární model pro testování hypotézy, zda $\mu_1 = \mu_2$.

Řešení. Budeme postupovat velmi podobně jako v odstavci 9.57 vedlejšího sloupce. Tentokrát můžeme zapsat oba vektory Y_i do jednoho sloupce pod sebe a budeme uvažovat model

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \sigma Z.$$

9.50. Bodové a intervalové odhady. Nyní máme vše připravené pro odhady hodnot parametrů v kontextu frekvenční statistiky. Budeme si postup ilustrovat na konkrétním jednoduchém příkladu. Řekněme, že máme v kurzu s 500 studenty výsledky jejich spokojenosti z ankety z minulého semestru ve formě bodů 1-10. Předpokládejme, že spokojenost jednotlivých studentů X_i je aproximována náhodnou veličinou s rozdělením $\mathrm{N}(\mu,\sigma^2)$, přičemž zjištěné hodnoty z celé populace minulého semestru jsou $\mu=6$, $\sigma=2$.

V běžícím semestru je provedeno namátkové šetření u 15 studentů, protože panuje obava, že nový vyučující má ještě výrazně horší ohlasy. Výsledkem je hodnocení, kde se vyskytují dvě 3, tři 4, tři 5, pět 6 a dvě 7. Výběrový průměr je tedy $\bar{X}=5{,}133$, výběrový rozptyl $S^2=1{,}695$.

Díky našim předpokladům víme, že $\bar{X}\sim \mathrm{N}(\mu,\sigma^2/n)$ a tedy $Z=\sqrt{n}\frac{\bar{X}-\mu}{\sigma}\sim \mathrm{N}(0,1)$. Pro vyjádření spolehlivosti našeho odhadu tedy můžeme počítat interval, který bude odhadovaný parametr obsahovat s předem zvolenou pravděpodobností $100(1-\alpha)\%$. Hovoříme přitom o hladině spolehlivosti $0<\alpha<1$. Nejprve považujme za neznámý nový parametr μ, zatímco o rozptylu budeme (ať už oprávněně nebo ne) předpokládat, že zůstal stejný. Dostaneme okamžitě

$$\begin{aligned} 1-\alpha = P(|Z|<z(\alpha/2)) &= P\left(\left|\sqrt{n}\frac{\bar{X}-\mu}{\sigma}\right|<z(\alpha/2)\right) \\ &= P\left(\bar{X}-\frac{\sigma}{\sqrt{n}}z(\alpha/2)<\mu<\bar{X}+\frac{\sigma}{\sqrt{n}}z(\alpha/2)\right) \end{aligned}$$

a našli jsme interval, jehož hranice jsou náhodné veličiny a který s předem zadanou pravděpodobností bude obsahovat odhadovaný parametr μ. Střed tohoto intervalu nazýváme *bodovým odhadem* pro parametr μ, celý interval pak *intervalovým odhadem*. Výsledek pak můžeme interpretovat i tak, že na hladině spolehlivosti α odhadovaný parametr μ je nebo není odlišný od jiné hodnoty μ_0.

V případě našich dat vyjdou např. pro hodnoty $\alpha=0{,}05$ a $\alpha=0{,}1$ intervaly

$$\mu\in(4{,}121,\ 6{,}145),\quad \mu\in(4{,}284,\ 5{,}983).$$

Na hladině spolehlivosti 5% tedy nemůžeme potvrdit, že se názor studentů na výuku nového učitele oproti minulému zhoršil, protože uvedený interval obsahuje i hodnotu $\mu_0=6$. Na úrovni 10% už takový úsudek uděláme, protože dřívější hodnota z minulého semestru $\mu_0=6$ už do našeho intervalu nepadne.

Pokud bychom ale předpokládali, že u jiného (horšího) učitele bude patrně i rozptyl odpovědí jiný (třeba se studenti více shodnou na špatném hodnocení), museli bychom postupovat trochu odlišně. Místo normalizované veličiny Z výše budeme stejně postupovat s veličinou

$$T=\sqrt{n}\frac{\bar{X}-\mu}{S}.$$

Jak jsme viděli, má tato náhodná veličina rozdělení pravděpodobnosti $T\sim \mathrm{t}_{n-1}$, kde v našem případě je $n=15$. Vyjde tak intervalový odhad

$$\bar{X}-\frac{S}{\sqrt{n}}t_{n-1}(\alpha/2)<\mu<\bar{X}+\frac{S}{\sqrt{n}}t_{n-1}(\alpha/2)$$

a po dosazení našich dat na úrovních $\alpha=0{,}05$ a $\alpha=0{,}03$ máme

$$\mu\in\big(4{,}412,\ 5{,}854\big),\quad \mu\in\big(4{,}321,\ 5{,}945\big),$$

Budeme pracovat s aritmetickými průměry jednotlivých vektorů $\bar{Y}_1$ a $\bar{Y}_2$. Přímá aplikace obecného vzorce z teorie dává odhad b ve tvaru

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} n_1+n_2 & n_2 \\ n_2 & n_2 \end{pmatrix}^{-1} \begin{pmatrix} n_1\bar{Y}_1 + n_2\bar{Y}_2 \\ n_2\bar{Y}_2 \end{pmatrix} =$$
$$= \frac{1}{n_1 n_2} \begin{pmatrix} n_2 & -n_2 \\ -n_2 & n_1+n_2 \end{pmatrix} \begin{pmatrix} n_1\bar{Y}_1 + n_2\bar{Y}_2 \\ n_2\bar{Y}_2 \end{pmatrix} = \begin{pmatrix} \bar{Y}_1 \\ \bar{Y}_2 - \bar{Y}_1 \end{pmatrix}$$

a matice $C = (X^T X)^{-1}$, kde X je dvousloupcová matice s nulami a jedničkami z našeho modelu, vychází

$$C = \begin{pmatrix} \frac{1}{n_1} & -\frac{1}{n_1} \\ -\frac{1}{n_1} & \frac{1}{n_1} + \frac{1}{n_2} \end{pmatrix}.$$

Testujeme tedy hypotézu $\mu_1 = \mu_2$, to znamená, že testujeme, zda je $\beta_2 = 0$. K tomu je proto vhodné použít statistiku

$$T = \frac{\bar{Y}_2 - \bar{Y}_1}{S} \left(\frac{n_1 n_2}{n_1 + n_2} \right)^{\frac{1}{2}},$$

kde za směrodatnou odchylku S dosazujeme

$$S^2 = \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2} (Y_{2i} - \bar{Y}_2)^2 \right).$$

Tato statistika má rozdělení $\mathrm{t}_{n_1+n_2-2}$ a nulovou hypotézu $\mu_1 = \mu_2$ proto zamítáme na hladině α, když platí

$$|T| \geq \mathrm{t}_{n_1+n_2-2}(\alpha). \qquad \square$$

9.85. V JZD[1] Tempo sledovali v pěti různých dnech dojivost krav a naměřili postupně tyto výsledky: 15, 14, 13, 16 a 17 hektolitrů. V JZD Boj, ve kterém mají stejný počet krav, měřili přibližně ve stejnou dobu, nicméně v sedmi různých dnech: 12, 16, 13, 15, 13, 11, 18 hektolitrů.

a) Určete 95% interval spolehlivosti pro dojivost krav v JZD Boj, a 95% interval spolehlivosti pro dojivost krav v JZD Tempo.

b) Na pětiprocentní hladině otestujte hypotézu, že v obou družstvech mají stejně kvalitní krávy.

Předpokládejte, že dojivost krav v jednotlivých dnech se řídí normálním rozdělením. Oba výpočty proveďte jak za předpokladu, že v družstvech mají k dispozici údaje z předchozích dlouhodobých měření, ve kterých byla směrodatná odchylka $\sigma = 2\ hl$ mléka, tak v případě, že údaje z předchozích měření nejsou k dispozici.

Řešení. Nejprve spočítejme výsledky za předpokladu známého rozptylu. K určení intervalu spolehlivosti použijeme statistiku

$$U = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

takže už na úrovni 3% spolehlivosti máme za to, že je názor na učitele skutečně horší. To odpovídá intuici, že nejspíš by výrazně menší výběrová směrodatná odchylka $S = 1,302$ než odchylka $\sigma = 2$ z minulého šetření také měla být podstatná pro naše úvahy.

9.51. Věrohodnost odhadů. Matematicky jsou intervalové a bodové odhady jednoduché a patrně dobře pochopitelné. Daleko horší je to s interpretací praktickou. Jednak je problematické ověřit všechny předpoklady o náhodnosti výběru, ale hlavně ve složitějších případech budeme mít problém s „věrohodností odhadů".

Jako matematici se praktickému problému nejlépe vyhneme tak, že podáme definici chybějícího pojmu. Obecně chceme pracovat s náhodným výběrem o rozsahu n. Implicitně stále předpokládáme, že jde o nezávislé náhodné veličiny X_i se shodným rozdělením pravděpodobnosti, které ale závisí na neznámém, obecně vektorovém, parametru θ.

Snažíme se najít nějakou *výběrovou statistiku* T, tj. funkci náhodných veličin $X_1, X_2, \ldots$, která v nějakém (matematickém) smyslu bude dobře odhadovat skutečnou hodnotu parametru θ. Říkáme, že je T *nestranným odhadem* parametru θ, jestliže je $\mathrm{E}\,T = \theta$. Střední hodnota $\mathrm{E}(T - \theta)$ se nazývá *vychýlení odhadu* T.

Často nás zajímá také asymptotické chování odhadu, tj. jak se chová při limitním přechodu $n \to \infty$. Říkáme, že je $T = T(n)$ *konzistentním odhadem* parametru θ, jestliže konverguje $T(n)$ v pravděpodobnosti k θ, tj. pro každé $\varepsilon > 0$

$$\lim_{n\to\infty} P(|T(n) - \theta| < \varepsilon) = 1.$$

Čebyševova nerovnost nám okamžitě dává

$$P(|T(n) - \mathrm{E}\,T_n| < \varepsilon) \geq 1 - \frac{\operatorname{var} T(n)}{\varepsilon^2}.$$

Pokud předpokládáme, že $\lim_{n\to\infty} \mathrm{E}\,T(n) = \theta$, pak zároveň pro dostatečně velká n platí

$$P(|T(n) - \theta| < 2\varepsilon) \geq P(|T(n) - \mathrm{E}\,T(n)| < \varepsilon) \geq 1 - \frac{\operatorname{var} T(n)}{\varepsilon^2}.$$

Dokázali jsme užitečné tvrzení:

Věta. *Předpokládejme, že platí* $\lim_{n\to\infty} \mathrm{E}\,T(n) = \theta$ *a zároveň předpokládejme* $\lim_{n\to\infty} \operatorname{var} T(n) = 0$. *Pak je* $T(n)$ *konzistentním odhadem pro* θ.

Jednoduchým příkladem pro použití této věty je rozptyl

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{n-1}{n} S^2.$$

Protože nestranným odhadem je podle věty z odstavce 9.48 S^2, víme, že $\hat{\sigma}^2$ nestranný odhad není. Zřejmě však platí $\lim_{n\to\infty} \hat{\sigma}^2 = \sigma^2$ a lze přímo spočíst také

$$\lim_{n\to\infty} \operatorname{var} \hat{\sigma}^2 = \lim_{n\to\infty} \operatorname{var} S^2 = \lim_{n\to\infty} \frac{2\sigma}{n-1} = 0.$$

Je tedy statistika s^2 konzistentním odhadem rozptylu.

[1] JZD — jednotné zemědělské družstvo — zemědělské družstvo vzniklé násilnou kolektivizací v padestátých letech dvacátého století.

která má standardizované normální rozdělení pravděpodobnosti (viz 9.26). Interval spolehlivosti pak je (viz 9.50)

$$\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z(\alpha/2),\, \bar{X} + \frac{\sigma}{\sqrt{n}} z(\alpha/2)\right),$$

kde $\alpha = 0,05$. Nyní pouze dosadíme číselné hodnoty. Pro údaje z JZD Tempo tak dostáváme výběrový průměr

$$\overline{X}_1 = \frac{15 + 14 + 13 + 16 + 17}{5} = 15,$$

z tabulek či matematického softwaru zjistíme, že $z(0{,}025) = 1{,}96$ a dostáváme interval

$$\left(15 - \frac{2}{\sqrt{5}} 1{,}96,\, 15 + \frac{2}{\sqrt{5}} 1{,}96\right) \doteq (13{,}25;\, 16{,}75).$$

Pro JZD Boj pak dostáváme

$$\overline{X}_2 = \frac{12 + 16 + 13 + 15 + 13 + 11 + 18}{7} = 14,$$

a 95% interval spolehlivosti pro hodnotu dojivosti krav v JZD Boj tak je

$$(12{,}52;\, 15{,}48).$$

Pokud je rozptyl měření neznámý, použijeme k jeho odhadu tzv. výběrový rozptyl a k určení intervalu spolehlivosti pak statistiku

$$T = \frac{\overline{X} - \mu}{S\sqrt{n}},$$

která má Studentovo rozložení pravděpodobnosti s $n-1$ stupni volnosti (viz též 9.50). Potom analogicky obdržíme 95% interval spolehlivosti

$$\left(\bar{X} - \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2),\, \bar{X} + \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2)\right).$$

Pro konkrétní hodnoty pak dostáváme pro JZD Tempo výběrový rozptyl

$$S_1^2 = \frac{0^2 + (-1)^2 + (-2)^2 + 1^2 + 2^2)}{4} = 2{,}5,$$

tedy $S \doteq 1{,}58$. Dále je $t_4(0,025) \doteq 2,78$. 95% interval spolehlivosti hodnot dojivosti krav v JZD Tempo tedy je

$$(13{,}03;\, 16{,}97).$$

Pro JZD Boj pak dostáváme výběrový rozptyl $S_2^2 = 6$ a hledaný interval spolehlivosti je pak

$$(11{,}73;\, 16{,}27).$$

b) Jestliže srovnáváme střední hodnoty dojivosti v obou družstvech, jedná se o porovnání středních hodnot dvou nezávislých výběrů z normálních rozložení. V případě neznámých rozptylů měření navíc předpokládejme, že rozptyl měření je v obou družstvech stejný.

Je vcelku zřejmé, že pro stejný parametr můžeme mít k dispozici spoustu nestranných odhadů. Např. jsme viděli, že aritmetický průměr $\bar{X}$ je nestranným odhadem střední hodnoty θ rozdělení veličin X_i. Samozřejmě je ale třeba hodnota X_1 také nestranným odhadem θ. Chceme proto najít nejlepší odhad T ve třídě uvažovaných statistik, které jsou nestrannými nebo konsistentními odhady. Zpravidla máme za to, že nejlepším odhadem je ten, který má ze všech uvažovaných nejmenší možný rozptyl. Připomeňme, že rozptyl vektorové statistiky T je dán kovarianční maticí, která bude, v případě nezávislých komponent, diagonální maticí s jednotlivými rozptyly komponent na diagonále. Nerovnostem mezi pozitivně definitními maticemi jsme již dříve dali jednoznačný smysl.

9.52. Maximální věrohodnost. Předpokládejme tedy, že náš výběr má komponenty s rozdělením, jehož hustota je dána funkcí $f(x, \theta)$ závislou na neznámém (obecně vektorovém) parametru θ. Sdružená hustota vektoru $(X_1, \ldots, X_n)$ je díky předpokládané nezávislosti dána součinem funkcí

$$f(x_1, \ldots, x_n, \theta) = f(x_1, \theta) \cdots f(x_n, \theta),$$

které říkáme *věrohodnostní funkce*.

Zajímáme se o takovou hodnotu $\hat{\theta}$, která maximalizuje na množině všech dostupných hodnot parametru věrohodnostní funkci. V diskrétním případě to znamená, že vybíráme takový parametr, při kterém vychází největší pravděpodobnost zjištěného výběru.

Zpravidla ale pracujeme s tzv. *logaritmickou věrohodnostní funkcí*

$$\ell(x_1, \ldots, x_n, \theta) = \ln f(x_1, \ldots, x_n, \theta) = \sum_{i=1}^{n} \ln f(x_i, \theta),$$

protože díky monotonnímu chování funkce ln je maximalizace věrohodnostní funkce ekvivalentní požadavku maximalizace logaritmické věrohodnostní funkce. Pokud je pro nějaké hodnoty $f(x_1, \ldots, x_n) = 0$, klademe $\ell(x_1, \ldots, x_n, \theta) = -\infty$.

V případě diskrétních náhodných veličin použijeme stejnou definici s pravděpodobnostní funkcí místo hustoty, tj.

$$\ell(x_1, \ldots, x_n, \theta) = \sum_{i=1}^{n} \ln(P(X_i = x_i|\theta)).$$

Princip je dobře vidět na náhodném výběru z normálního rozdělení $\mathrm{N}(\mu, \sigma^2)$ o rozsahu n. Neznámé parametry jsou μ nebo σ, nebo oba. Uvažovaná hustota je

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

a tedy logaritmováním okamžitě vidíme

$$\ell(x, \mu, \sigma) = -n\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Vyšetřeme tedy hypotézu za předpokladů známých, uvedených rozptylů $\sigma_1^2 = \sigma_2^2 = 4$. Použijeme statistiku

$$\begin{aligned} U &= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \\ &= \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1), \end{aligned}$$

kde μ_1 a μ_2 jsou neznámé střední hodnoty dojivosti ve zkoumaných družstvech a n_1, n_2 jsou počty měření. Tato statistika má, jak naznačeno, standardizované normální rozdělení. Hypotézu na 5% hladině zamítneme, právě když absolutní hodnota statistiky U bude větší než $z_{0{,}025}$, neboli právě když 0 nebude ležet v 95% intervalu spolehlivosti pro rozdíl středních hodnot dojivosti v jednotlivých družstvech. Po dosazení číselných hodnot dostáváme

$$U = \frac{15 - 14}{\sqrt{\frac{4}{5} + \frac{4}{7}}} \doteq 0{,}854.$$

Je tedy $|U| < z(0{,}025) = 1{,}96$ a hypotézu o rovnosti středních hodnot dojivosti v obou družstvech na 5% hladině nezamítáme. Dosažená p-hodnota testu (viz 9.55) je 39, 4%, tudíž jsme se k zamítnutí hypotézy moc nepřiblížili (pravděpodobnost, že hodnota zkoumané statistiky bude menší než 0,854 je při platnosti nulové hypotézy 60, 6%).

Pokud neznáme rozptyly měření, ale víme, že v obou družstvech musí být stejné, použijeme statistiku

$$\begin{aligned} K &= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \\ &= \frac{\overline{X}_1 - \overline{X}_2}{S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}, \end{aligned}$$

kde

$$S_* = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

Po dosazení číselných hodnot dostáváme $K \doteq 0{,}796$, $|K| < t_{10}(0, 025) = 2{,}2281$, nulovou hypotézu tedy opět nezamítáme. Dosažená p-hodnota testu je 44,6%, tedy ještě větší než v testu předešlém. □

9.86. Analýza rozptylu jednoduchého třídění. Pro $k \geq 2$ nezávislých výběrů Y_i o rozsahu n_i z normálních rozdělení se stejným rozptylem použijte lineární model na testování hypotézy, že všechny střední hodnoty jednotlivých výběrů jsou shodné.

Maximum najdeme pomocí derivací (všimněme si, že σ^2 chápeme při derivování jako symbol pro proměnnou):

$$\begin{aligned} \frac{\partial \ell}{\partial \mu} &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (-2)(x_i - \mu) = \frac{1}{\sigma^2}\Big(-n\mu + \sum_{i=1}^{n} x_i\Big) \\ \frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = \\ &= \frac{1}{2\sigma^4}\Big(-n\sigma^2 + \sum_{i=1}^{n} (x_i - \mu)^2\Big). \end{aligned}$$

Vidíme tedy, že jediné kritické body jsou dány právě volbou $\hat{\mu} = \bar{X}$ a $\hat{\sigma}^2 = s^2$. Dosazením těchto hodnot do matice druhých derivací dostaneme Hessián funkce ℓ

$$\begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2(\hat{\sigma}^2)^2} \end{pmatrix}.$$

Je tedy vidět že jde skutečně o dosažené maximum a protože je jediné, musí jít o globální maximum. Ověřili jsme tedy, že střední a hodnota a rozptyl jsou skutečně maximálně věrohodné odhady pro μ a σ, tak jak jsme je používali výše.

9.53. Bayesovské odhady. Vraťme se teď k příkladu z odstavce 9.50 z pohledu Bayesovské statistiky. Úplně tedy otáčíme náš přístup a zjištěná data $X_1, \ldots, X_{15}$, tj. body vyjadřující spokojenost dotázaných studentů na škále $1, \ldots, 10$ bodů, budeme chápat jako konstanty. Naopak, odhadovaný parametr μ, tj. střední hodnota bodů vyjadřujících spokojenost, bude náhodnou veličinou, jejíž rozložení chceme odhadnout.

Za tímto účelem zkusme interpretovat Bayesův vzorec pro podmíněnou pravděpodobnost na úrovni pravděpodobnostních funkcí, resp. hustot pravděpodobností, následujícím způsobem. Má-li vektor (X, Θ) sdruženou hustotu $f(x, \theta)$, pak podmíněná pravděpodobnost komponenty Θ za podmínky $X = x$ je dána hustotou

$$g(\theta|x) = \frac{f(x|\theta)g(\theta)}{f(x)},$$

kde $f(x)$ a $g(\theta)$ jsou marginální hustoty pravděpodobností.

Jestliže tedy máme dánu *apriorní* hustotu rozložení pravděpodobnosti $g(\theta)$ odhadovaného parametru θ a známe také hustotu pravděpodobnosti $f(x|\theta)$, můžeme ze vztahu spočíst *aposteriorní* hustotu pravděpodobnosti $g(\theta|x)$ vycházející právě ze zjištěných dat. Protože data X jsou přitom konstantní, nepotřebujeme ve skutečnosti vůbec počítat s hodnotou $f(x)$ a při úvahách budeme pracovat jen „až na konstantní násobek“. Ten je totiž stejně určen na konci úvah jednoznačně požadavkem, aby vyšla dobře definovaná hustota rozdělení pravděpodobnosti $g(\theta|x)$. Budeme pro tento účel používat zápis $Q \propto R$, jestliže existuje konstanta C taková, že pro výrazy Q a R platí $Q = CR$.

Abychom byli technicky co nejblíže k úvahám v odstavci 9.50, budeme pracovat s normálními rozděleními $\mathrm{N}(\mu, \sigma^2)$. Předpokládejme, že na univerzitě je spokojenost studentů v jednotlivých předmětech náhodná veličina $X \sim \mathrm{N}(\theta, \sigma^2)$, zatímco parametr θ dosahovaný jednotlivými učiteli je náhodná veličina $\theta \sim \mathrm{N}(a, b)$.

Řešení. Postup je zde velice podobný minulému příkladu, platnost testované hypotézy je ale ekvivalentní tvrzení, že platí podmodel, ve kterém mají všechny složky náhodného vektoru Y vzniklého sloučením daných k vektorů Y_i stejnou střední hodnotu.

Použitý model tedy bude mít tvar

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{k1} \\ \vdots \\ Y_{kn_k} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} + \sigma Z.$$

Snadno spočteme odhady středních hodnot μ_i pomocí aritmetických průměrů

$$\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}.$$

Odtud dostaneme odhad $\hat{Y}_{ij} = \bar{Y}_i$ a proto dostaneme reziduální součet čtverců ve tvaru

$$RSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2.$$

Odhadem společné střední hodnoty v uvažovaném podmodelu je

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n} \sum_{i=1}^{k} n_i \bar{Y}_i,$$

kde $n = n_1 + \cdots + n_k$, a reziduální součet čtverců v tomto podmodelu je

$$RSS^0 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2.$$

V původním modelu máme k nezávislých parametrů μ_i, zatímco v podmodelu zůstal jediný parametr μ, testovaná statistika má proto tvar

$$F = \frac{(n-k)}{(k-1)} \frac{(RSS^0 - RSS)}{RSS}.$$

□

J. Lineární regrese

S lineární regresí jsme se už setkali ve třetí kapitole, v odstavci ‖3.46‖. Nyní se stejný princip budeme snažit využít k vyřešení problémů, které bývají studovány statistiky.

Standardním příkladem užití lineární regrese je „proložení přímky“ danými daty. Máme tedy posloupnost měření, ve kterých zaznamenáváme hodnoty dvou veličin u nichž předpokládáme lineární závislost. Klasickým příkladem je závislost výšky syna na výšce otce.

Můžeme tedy počítat (pořád až na konstantní násobky, tj. ignorujeme součinitele, ve kterých nevystupuje ani x ani θ)

$$\begin{aligned} g(\theta|x) &\propto f(x|\theta) g(\theta) \\ &\propto \exp\left(-\frac{(x-\theta)^2}{2\sigma^2} - \frac{(\theta-a)^2}{2b^2} \right) \\ &\propto \exp\left(-\frac{1}{2} \left(\theta^2 \left(\frac{1}{\sigma^2} + \frac{1}{b^2} \right) - 2\theta \left(\frac{x}{\sigma^2} + \frac{a}{b^2} \right) \right) \right) \\ &\propto \exp\left(-\frac{1}{2} \left(\theta - \frac{b^2 x + \sigma^2 a}{\sigma^2 b^2} \frac{\sigma^2 b^2}{b^2 + \sigma^2} \right)^2 \left(\frac{b^2 \sigma^2}{b^2 + \sigma^2} \right)^{-1} \right). \end{aligned}$$

Tím jsme ale už ukázali, že hledané rozložení pro θ je

$$\theta \sim \mathrm{N}\left(\frac{b^2}{b^2 + \sigma^2} x + \frac{\sigma^2}{b^2 + \sigma^2} a, \frac{b^2 \sigma^2}{b^2 + \sigma^2} \right).$$

Tento výsledek bychom mohli interpretovat např. tak, že když z dlouhodobého vyhodnocování anket známe parametry a, b, σ, můžeme po vyjádření nějakého studenta upřesnit apriorní představu o parametrech pro jeden konkrétní předmět. Ve výsledném odhadu rozložení je pak střední hodnota dána váženým průměrem zjištěné hodnoty x a apriorně předpokládané střední hodnoty a, v závislosti na rozptylech σ a b.

9.54. Interpretace v Bayesovské statistice. Zkusíme teď porozumět úvahám z předchozího odstavce ve srovnání s frekventistickou interpretací z 9.50. Asi namítneme, že jediný dotaz těžko má tolik ovlivnit náš názor.

Ve skutečnosti ale pro $\sigma \to 0$ je váha jednoho názoru stále rostoucí a v našem výsledku tomu odpovídá 100% váha u x v případě $\sigma = 0$. Je to plně v souladu s interpretací, že Bayesovská statistika je pravděpodobnostní rozšíření standardní diskrétní matematické logiky. Jestliže máme rozptyl σ prakticky nulový, pak je tedy v tomto smyslu skoro jisté, že názor kteréhokoliv studenta je naprosto vypovídající o celé populaci.

Ve skutečnosti jsme v odstavci 9.50 pracovali s výběrovým průměrem $\bar{X}$ výsledku šetření. Ten můžeme použít i v předchozím výpočtu, protože jde opět o normální rozdělení, jen budeme místo σ^2 dosazovat σ^2/n. Pro zjednodušení zápisu si definujme konstantu

$$c_n = \frac{nb^2}{nb^2 + \sigma^2}$$

a aposteriorní odhad pro θ na základě zjištěni výběrového průměru $\bar{X}$ má rozložení s parametry

$$\theta \sim \mathrm{N}(c_n \bar{X} + (1 - c_n)a, c_n \sigma^2/n).$$

Jak se dalo očekávat, pro rostoucí n se bude střední hodnota našeho rozdělení pro θ stále více blížit výběrovému průměru a jeho rozptyl půjde k nule. Čím je tedy n větší, tím více se blížíme bodovému odhadu z frekventistického přístupu.

Přínosem Bayesovského přístupu je, že s použitím odhadnutého rozdělení můžeme odpovídat na dotazy typu „s jakou pravděpodobností je nový vyučující horší než předchozí?“ Použijeme stejná data jako v 9.50 a přidáme potřebné apriorní údaje. Předpokládejme, že máme docela dobře hodnocené učitele (protože by asi jinak na škole nevydrželi), takže uvažujme pro určitost $a = 7{,}5$, $b = 2{,}5$ a a ponecháváme směrodatnou odchylku $\sigma = 2$.

9.87. Určete lineární regresní model pro závislost veličiny Y na veličině X na základě naměřených seznamů dat: $X = [1, 4, 5, 7, 10]$, $Y = [3, 7, 8, 12, 18]$.

Řešení. K určení parametrů regresní přímky použije vztahů odvozených v 9.57. Podle metody nejmenších čtverců se snažíme se minimalizovat vzdálenost vektoru $b_1X + b_0$ od vektoru Y v závislosti na parametrech b_1 a b_0. Tato vzdálenost, jak víme například ze druhé kapitoly, je minimální pro kolmý průmět vektoru Y do vektorového podprostoru generovaného vektory $(1, \ldots, 1)$ a $(x_1, \ldots, x_n)$. Pro parametry b_0, b_1 regresní přímky $Y = b_1X + b_0$ tak dostáváme

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} =$$
$$= \frac{(1-5{,}4)(3-9{,}6) + \cdots + (10-5{,}4)(18-9{,}6)}{((1-5{,}4)^2 + (4-5{,}4)^2 + (5-5{,}4)^2 + (7-5{,}4)^2 + (10-5{,}4)^2} =$$
$$= 1{,}677.$$

Teď již snadno dopočteme i koeficient b_0:

$$b_0 = \bar{Y} - b_1\bar{x} = 0{,}5442.$$

Hledaná lineární závislost je tedy

$$Y = 1{,}677 \cdot X + 0{,}5442.$$

Poznamenejme, že v tomto modelu hrají veličiny X a Y naprosto rovnocennou úlohu. Stejnou metodou jsme mohli získat závislost X na Y:

$$X = 0{,}5867 \cdot Y - 0{,}2322.$$

□

Poznámka. Rozmyslete si, proč lineární regresní model závislosti X na Y nelze získat pouhým vyjádřením X z lineárního regresního modelu závislosti Y na X.

Poznámka. V řadě reálných situací je závislost veličin jasně dána, například je-li jednou z veličin čas.

9.88. Orbitální stanice naměřila v pěti po sobě jdoucích dnech, ve stejnou hodinu následující rychlosti neznámého vesmírného tělesa (v km/s): 10, 11,4, 13,1, 15,8 a 18,7. Odhadněte rychlost tělesa desátého dne.

Řešení. Zde je vhodné si všimnout, že rychlost se v čase „od pohledu“ nemění lineárně (nárůsty rychlosti se neustále zvyšují). Lze tedy vyslovit domněnku, že je těleso přitahováno gravitační silou k nějakému jinému tělesu. Potom by jeho rychlost byla kvadratickou funkcí času. Zkusme tedy metodou nejmenších čtverců proložit co nejpřesněji kvadratickou funkci danými daty. Postup je stejný, jako

Měli jsme $n = 15$ a výběrový průměr 5,133. Dosazením dostaneme aposteriorní odhad pro rozdělení

$$\theta \sim \mathrm{N}(5{,}230,\ 0{,}256).$$

Zajímá nás teď $P(\theta < 6)$. Odpověď získáme dotazem na hodnotu distribuční funkce příslušného normálního rozdělení pro argument 6 (umí to i excel). Dostaneme odpověď přibližně 93, 6%. Je tedy podobná, jako jsme viděli v odstavci 9.50 v případě předpokladu o konstantním známém rozptylu.

Všimněme si, jak se tady projevuje apriorní předpoklad o rozložení parametru θ u všech učitelů, tj. jistá míra naší důvěry, že by měli být učitelé spíše lepší. Kdyby měl statistik důvod předpokládat, že pro konkrétně poptávaného učitele je skutečná střední hodnota a posunutá, řekněme na $a = 6$ stejně jako u ankety minulého učitele (např. protože jde o těžký a neoblíbený předmět), pak bychom dostali pravděpodobnost, že je jeho skutečný parametr menší než 6, přibližně 95,0% (pokud bychom ale za viditelně horší považovali až střední hodnotu menší než 5,5, pak už to bude jen přibližně 75%). V případě dosazení $a = 5$ to již bude 96,8%. Stejně tak hraje roli i rozptyl b^2. Např. apriorní odhad $a = 6$, $b = 3{,}5$ vede na pravděpodobnost 95,2%.

Právě jsme se mimoděk dotkli jiného velice podstatného bodu a to *analýzy citlivosti*. Jistě bychom rádi pracovali ve výše uvedeném příkladu s modelem, kde malá změna apriorního předpokladu bude mít jen malý vliv na aposteriorní výsledek. Zdá se, že v našem případě to tak skutečně je, nepůjdeme tu do detailních úvah.

Úplně stejný model s exponenciálními rozděleními se prakticky používá při posouzení relevance výstupu z IQ testu u jedné osoby (nebo jiné obdobné zkoušky, kde lze očekávat dobré přiblížení rozdělení pravděpodobnosti výsledků v populaci pomocí normálního rozdělení), o které máme apriorní předpoklad, do jaké skupiny by měla patřit. Jiným dobrým příkladem (ovšem s jinými rozloženími) mohou být praktické úlohy z pojišťovnictví, kde je účelné odhadovat parametry tak, aby byly zahrnuty jak vlivy experimentu na konkrétní položce, tak celková očekávání přes populaci.

9.55. Poznámky o testování hypotéz. Vrátíme se zpět k možnostem rozhodování, zda nějaký jev nastal či nenastal v kontextu frekvenční statistiky. Opřeme se o postup v intervalových odhadech výše.

Uvažujme tedy nějaký náhodný vektor $X = (X_1, \ldots, X_n)$ (vzniklý z náhodného výběru) se sdruženou distribuční funkcí $F_X(x)$. Za *hypotézu* budeme považovat nějaké tvrzení o rozdělení určeném touto distribuční funkcí. Zpravidla přitom formulujeme dvě hypotézy H_0 a H_A, kde první se tradičně říká *nulová hypotéza* a druhé *alternativní hypotéza*. Výsledkem testu je rozhodnutí založené na konkrétní realizaci náhodného vektoru X (hovoříme také o *testu*), zda hypotézu H_0 zamítnout nebo nezamítnout ve prospěch hypotézy H_A.

Vznikají nám přitom možné chyby dvou typů. *Chyba prvního druhu* nastává, když zamítneme H_0, přestože je platná. *Chyba druhého druhu* nastane, když naopak nezamítneme H_0, ačkoliv platná není. Rozhodování frekventistického statistika probíhá tak, že si vybere tzv. *kritický obor* W, tj. množinu výsledků realizace testu, při kterých hypotézu zamítá. Velikost kritického oboru přitom volí tak, aby platnou hypotézu zamítal s pravděpodobností nejvýše α. To znamená, že požadujeme předem dané ohraničení pravděpodobnosti chyby prvního druhu tzv. *hladinou testu* α. Často se volí

kdybychom prováděli lineární regresi vektoru $v = (v_1, v_2, \ldots, v_n)$ závislého na vektoru $x = (x_1, \ldots, x_n)$ a vektoru $x^2 = (x_1^2, \ldots, x_n^2)$. Této metodě se říká *kvadratická regrese*. Hledáme tedy vektor parametrů $b = (b_0,\ b_1,\ b_2)$ tak aby veličina $b_2x^2 + b_1x + b_0$ odhadovala y. Sestavme tedy matici X hodnot nezávislých proměnných:

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 4 & 25 \end{pmatrix},$$

a vektor parametrů $b = (b_0, b_1, b_2)$ dopočítáme podle (9.18):

$$b = (X^T X)^{-1} X^T v \doteq (9{,}26;\ 0{,}47;\ 0{,}29).$$

Hledaný kvadratický odhad je potom

$$v = 0{,}29x^2 + 0{,}47x + 9{,}26,$$

Odhadovaná rychlost desátého dne je tedy přibližně 42,96 km/s. V modelu klasické lineární regrese bychom dostali přiblížení

$$v = 2{,}18x + 7{,}26,$$

což by pro rychlost desátého dne dávalo 29,06 km/s. Rozdíl v odhadech je značný. Ukazuje to, že analýza situace je velmi významnou součástí statistiky. □

K. Bayesovská analýza dat

9.89. Mějme Bernoulliův proces definovaný náhodnou veličinou $X \sim \mathrm{Bi}(n, \theta)$ s binomickým rozdělením pravděpodobnosti a předpokládejme, že parametr θ je přitom náhodnou veličinou s rovnoměrným rozdělením pravděpodobnosti na intervalu $(0, 1)$. Definujme *šanci na úspěch* v našem procesu jako veličinu $\gamma = \frac{\theta}{1-\theta}$. Jakou hustotu rozdělení má veličina γ?

Řešení. Intuitivně asi cítíme, že nepůjde o rovnoměrné rozdělení.

Označíme hledanou hustotu pravděpodobnosti $f(s)$ a ze vztahu mezi θ a γ spočteme $\theta = \frac{\gamma}{1+\gamma}$. Také okamžitě vidíme, že hustota pravděpodobnosti veličiny γ bude nenulová pouze pro kladné hodnoty proměnné. Zadání můžeme nyní zformulovat tak, že požadujeme

$$\Theta = P(\theta < \Theta) = P\Big(\gamma < \frac{\Gamma}{1+\Gamma}\Big) = \int_0^\Gamma f(s)ds, \tag{9.4}$$

kde $\Gamma = \frac{\Theta}{1-\Theta}$. Na pravé straně máme ovšem v horní mezi právě měnící se ohraničení γ a dostáváme tedy definiční vztah pro $f(s)$

$$f(s) = \left(\frac{s}{s+1}\right)' = \frac{1}{(s+1)^2}.$$

Hledaná hustota skutečně dává daleko větší pravděpodobnost malým hodnotám šance než velkým. □

hladiny testů $\alpha = 0{,}05$ nebo $\alpha = 0{,}01$. Prakticky užitečný je také postup, kdy určíme nejnižší možnou hladinu p testu, při které ještě hypotézu zamítáme a mluvíme pak o *dosažené hladině testu*, resp. *p–hodnotě testu*.

Zbývá tedy najít rozumný postup, jak volit kritické obory. Jistě to budeme chtít dělat tak, abychom zároveň co nejvíce omezili výskyt chyby druhého druhu. Zpravidla se k tomu čelu hodí věrohodnostní funkce $f(x, \theta)$, kterou jsme přiřadili náhodnému vektoru X již v odstavci 9.52. Pro jednoduchost předpokládejme, že máme jednorozměrný parametr θ a nulovou hypotézu formulujme tak, že rozdělení X je dáno funkcí $f(x, \theta_0)$, zatímco alternativní hypotéza je dána rozdělením $f(x, \theta_1)$ pro dvě různé konkrétní hodnoty θ_0 a θ_1. Naše představy o zamítání nebo přijímání hypotéz napovídají, že po dosazení hodnot konkrétního pokusu do věrohodnostní funkce budeme chtít hypotézu spíše přijímat, je-li $f(x, \theta_0)$ výrazně větší než $f(x, \theta_1)$.

Nabízí se tedy pro každou konstantu $c > 0$ uvažovat kritický obor

$$W_c = \{x;\ f(x, \theta_1) \geq c f(x, \theta_0)\}.$$

Když si vybereme zamýšlenou hladinu testu, budeme chtít zvolit takové c, aby platilo

$$\int_{W_c} f(x, \theta_0) = \alpha.$$

Tím máme zajištěno, že pro výsledek testu $x \in W_c$ při platnosti H_0 se skutečně dopustíme maximálně předepsané chyby prvního druhu. To ale můžeme zajistit i jinými kritickými obory W splňujícími také

$$\int_{W} f(x, \theta_0) = \alpha.$$

Zajímá nás ale také pravděpodobnost chyby druhého druhu, zkusíme si tedy odhadnout rozdíl

$$D = \int_{W_c} f(x, \theta_1) - \int_{W} f(x, \theta_1).$$

Oblasti, přes které integrujeme můžeme v obou případech rozdělit na společnou část $W \cap W_c$ a zbývající množinový rozdíl. Příspěvky společné části se přitom odečtou a zbude

$$D = \int_{W_c \setminus W} f(x, \theta_1) - \int_{W \setminus W_c} f(x, \theta_1).$$

Díky naší definici kritického oboru W_c ale nyní můžeme snadno odhadnout (opět přidáváme zpět stejné integrály přes společnou část množin)

$$\begin{aligned} D &\geq c \int_{W_c \setminus W} f(x, \theta_0) - c \int_{W \setminus W_c} f(x, \theta_0) = \\ &= c \int_{W_c} f(x, \theta_0) - c \int_{W} f(x, \theta_0) = c\alpha - c\alpha = 0. \end{aligned}$$

Tím jsme odvodili důležité tvrzení, tzv. *Neymanovo–Pearsonovo lemma*, že za výše uvedených předpokladů je W_c optimální kritický obor, který na předepsané úrovni minimalizuje chybu druhého druhu.

Intervalový odhad, jak jsme jej ilustrovali na příkladu v odstavci 9.50, je speciálním případem testování hypotéz, kdy H_0 má formu „střední hodnota spokojenosti studentů zůstala μ_0“, zatímco H_A říká, že bude rovna nějaké jiné hodnotě μ_1. Uvidíme za chvilku, že předchozí obecný

V odstavci 9.53 jsme viděli, že jestliže pracujeme v bayesovském přístupu s binomickým modelem rozdělení pravděpodobnosti náhodné veličiny $X \sim \mathrm{Bi}(n,\theta)$, bude nás zajímat její pravděpodobnostní funkce $f_X(k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$. Na tuto funkci se ale můžeme také dívat jako na podmíněnou pravděpodobnost $P(\theta|X=k)$ při apriorním rovnoměrném rozdělení pravděpodobnosti veličiny θ na intervalu $(0,1)$. Je to tedy právě aposteriorní rozdělení pravděpodobnosti veličiny θ odpovídající výsledku pokusu $X=k$. Následující příklad se týká obecné třídy takovýchto rozdělení pravděpodobnosti.

9.90. Najděte základní charakteristiky tzv. *Beta rozdělení* $\beta(a,b)$ s hustotou pravděpodobnosti tvaru

$$f_Y = \begin{cases} C\,y^{a-1}(1-y)^{b-1} & y \in (0,1) \\ 0 & \text{jinak.} \end{cases}$$

Řešení. Konstantu C je třeba volit jako reciprokou hodnotu integrálu $\int_0^1 y^{a-1}(1-y)^{b-1}dy$, což je funkce $\mathrm{B}(a,b)$, v matematické analýze (ale také technických vědách či fyzice) známá pod názvem *Beta funkce*. Když už známe funkci Gama, která zobecňuje diskrétní hodnoty faktoriálů, vyskočí na nás např. při následujícím výpočtu:

$$\begin{aligned}
\Gamma(x)\Gamma(y) &= \int_0^\infty \mathrm{e}^{-t}\,t^{x-1}\,dt \cdot \int_0^\infty \mathrm{e}^{-s}\,s^{y-1}\,ds = \\
&= \int_0^\infty \int_0^\infty \mathrm{e}^{-t-s}\,t^{x-1}s^{y-1}\,dt\,ds = \\
&\quad \text{(substituce } t = rq,\ s = r(1-q)\text{)} \\
&= \int_{r=0}^\infty \int_{q=0}^1 \mathrm{e}^{-r}(rq)^{x-1}\big(r(1-q)\big)^{y-1} r\,dq\,dr = \\
&= \int_{r=0}^\infty \mathrm{e}^{-r}\,r^{x+y-1}\,dr \cdot \int_{t=0}^1 q^{x-1}(1-q)^{y-1}dq = \\
&= \Gamma(x+y)\mathrm{B}(x,y).
\end{aligned}$$

dostáváme tedy obecný vztah

$$\mathrm{B}(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

a z vlastností Gamma funkce již snadno plyne, že pro přirozená kladná a, b bude platit

$$\mathrm{B}(n-k+1,k+1) = \frac{k!(n-k)!}{(n+1)!} = \frac{1}{n+1}\binom{n}{k}^{-1}.$$

Přímým výpočtem vidíme, že střední hodnota veličiny $X \sim \beta(a,b)$ s beta rozdělením je (využíváme vztah $\Gamma(z+1) = z\Gamma(z)$)

$$\mathrm{E}\,X = \frac{\mathrm{B}(a+1,b)}{\mathrm{B}(a,b)} = \frac{a}{a+b}.$$

Je-li $a=b$ vyjde střední hodnota i medián $\frac{1}{2}$.

postup povede v tomto případě na kritický obor zadaný požadavkem

$$|Z| = \left|\frac{\bar X - \mu_0}{\sigma}\sqrt{n}\right| \geq z(\alpha/2).$$

Všimněme si, že v definici kritického oboru není skutečná hodnota μ_1 podstatná a zformalizovali jsme proto v kontextu klasické pravděpodobnosti úlohu rozhodnout na předepsané hladině α, zda se střední hodnota μ změnila.

Jestliže ale chceme testovat z nějakého důvodu pouze, zda spokojenost poklesla, pak musíme předem předpokládat, že $\mu_1 < \mu_0$. Rozeberme si tento případ podrobněji. Kritický obor z Neymanova–Pearsonova lemmatu je určen nerovností

$$\frac{f(x,\mu_1,\sigma^2)}{f(x,\mu_0,\sigma)^2} = \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^n\left((x_i-\mu_1)^2-(x_i-\mu_0)^2\right)} \geq c.$$

Logaritmováním dostaneme, po drobné úpravě,

$$2\bar x(\mu_1-\mu_0) - (\mu_1^2-\mu_0^2) \geq \frac{2\sigma^2}{n}\ln c$$

a protože předpokládáme $\mu_1 < \mu_0$, dostáváme konečně

$$\bar x \leq \frac{\mu_1+\mu_0}{2} + \frac{\sigma^2}{n(\mu_1-\mu_0)}\ln c = y.$$

Konstantu c, tj. také rozhodující parametr y, přitom pro zvolenou hladinu α máme určenu tak, aby za předpokladu platnosti hypotézy H_0 platilo

$$\alpha = P(\bar X \leq y) = P\Big(\frac{\bar X-\mu_0}{\sigma}\sqrt{n} \leq \frac{y-\mu_0}{\sigma}\sqrt{n}\Big).$$

Díky předpokladu o platnosti hypotézy H_0 je veličina

$$Z = \frac{\bar X-\mu_0}{\sigma}\sqrt{n} \sim \mathrm{N}(0,1)$$

a proto znamená náš požadavek volbu $Z \leq -z(\alpha)$, která jednoznačně určuje optimální W_c.

Všimněme si, že tento kritický obor skutečně nezávisí na volbě hodnoty μ_1 a skutečnou hodnotu pro y jsem vůbec nepotřebovali vyjádřit. Podstatný byl pouze předpoklad $\mu_1 < \mu_0$.

V našem ilustračním příkladu v odstavci 9.50 tedy máme H_0 : $\mu = 6$ a alternativní hypotéza je H_A : $\mu < 6$. Rozptyl je $\sigma^2 = 4$. Test s $n = 15$ nám dal $\bar x = 5{,}133$. Dosazením dostaneme hodnotu $z = \frac{5{,}133-6}{4}\sqrt{15} = -1{,}678$ zatímco $-z(0{,}05) = -1{,}645$.

Hypotézu tedy na hladině 5% zamítáme a usuzujeme, že skutečně došlo ke zhoršení názoru studentů.

Když si za kritický obor zvolíme sjednocení oborů pro případy $\mu_1 < \mu_0$ a $\mu_1 > \mu_0$, dostaneme právě výsledek shodný s intervalovým odhadem, jak bylo zmíněno výše.

Poznamenejme závěrem, že v bayesovském přístupu je také možné přijímat nebo zamítat hypotézy víceméně v přímé vazbě na aposteriorní pravděpodobnosti jevů, jak jsme do jisté míry naznačili v odstavci 9.54 při interpretaci našeho konkrétního příkladu.

9.56. Lineární modely. Jako obvykle v analýze matematických problémů buď vystačíme s afinními závislostmi nebo skutečné vztahy jejich pomocí aproximujeme. Stejně tak ve statistice patří mnoho metod do tzv. lineárních modelů. Probereme si stručně jeden případ z frekvenční statistiky.

Přímo se také spočte rozptyl

$$\operatorname{var} X = \mathrm{E}(X - \mathrm{E}\,X)^2 = \frac{ab}{(a+b)^2(a+b+1)}.$$

Pro $a = b$ tedy dostáváme $\operatorname{var} X = \frac{1}{8a+4}$, což ukazuje, že pro rostoucí $a = b$ klesá rozptyl. V případě $a = b = 1$ dostáváme obyčejné rovnoměrné rozdělení na intervalu $(0, 1)$. □

9.91. V situaci stejné jako v předposledním příkladu předpokládejme, že v Bernoulliho procesu je šance zdaru θ náhodná veličina s rozdělením pravděpodobnosti $\beta(a, b)$. Jak bude vypadat rozdělení pravděpodobnosti veličiny $\gamma = \frac{\theta}{1-\theta}$? V čem bude zvláštní při $a = b = p$?

Řešení. V již řešeném příkladě jsme měli speciální případ s rovnoměrným rozdělením $\beta(1, 1)$. Můžeme tedy pokračovat v řešení v rovnosti $\|\|9.4\|\|$, kdy jsme tvar tohoto rozdělení použili. Dostáváme nyní na levé straně místo Θ výraz

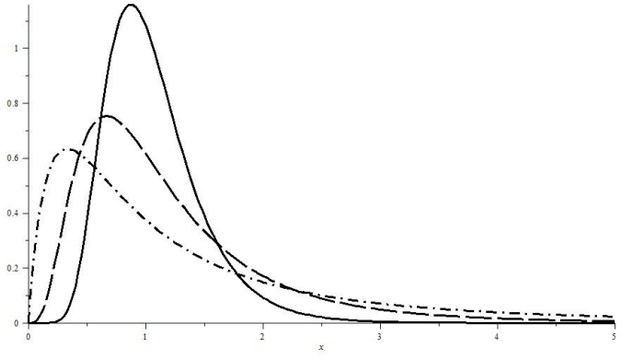
$$\frac{1}{\mathrm{B}(a, b)} \int_0^{\Theta} t^{a-1}(1-t)^{b-1} dt$$

a při derivování musíme použít pravidlo pro derivování integrálu s proměnnou horní mezí. Dostáváme proto pro hledanou hustotu

$$\mathrm{B}(a, b) f(s) = \left(\frac{s}{s+1}\right)^{a-1} \left(1 - \frac{s}{s+1}\right)^{b-1} \frac{1}{(s+1)^2} =$$

$$= \left(\frac{s^{a-1}}{s+1}\right)^{a+b}.$$

Na obrázku jsou vyneseny hustoty pro hodnoty $a = b = p = 2, 5, 15$.



Vidíme, že se naplňuje představa, že stejné a nepříliš malé hodnoty $a = b = p$ odpovídají nejvíce pravděpodobné hodnotě $\theta = \frac{1}{2}$, proto je hustota šance největší v okolí jedničky. Čím větší p, tím menší je rozptyl této veličiny. □

Budeme uvažovat náhodný vektor $Y = (Y_1, \ldots, Y_n)^T$ a budeme předpokládat, že platí

$$Y = X \cdot \beta + \sigma Z,$$

kde $X = (x_{ij})$ je konstantní matice reálných čísel s n řádky a $k < n$ sloupci a hodností k, β je neznámý konstantní vektor k parametrů modelu, Z je náhodný vektor, jehož n komponent má rozdělení $\mathrm{N}(0, 1)$, a $\sigma > 0$ je neznámý kladný parametr modelu. Hovoříme o *lineárním modelu* s úplnou hodností.

V praktických problémech jde často o to, že známe veličiny x_{ij} a snažíme se odhadnout nebo predikovat hodnotu Y. Například x_{ij} může vyjadřovat hodnocení i–tého studenta v j–tém semestru ($j = 1, 2, 3$) z matematiky a chceme vědět, jak tento student asi dopadne ve čtvrtém semestru. K tomu potřebujeme znát vektor β. Ten odhadneme na základě úplných pozorování, tj. ze znalosti Y (např. z výsledků v předchozích letech).

K odhadu vektoru β se často používá *metoda nejmenších čtverců*. To znamená, že chceme najít odhad $b \in \mathbb{R}^k$ tak, aby vektor $\hat{Y} = Xb$ minimalizoval druhou mocninu délky vektoru $Y - X\beta$.

To je ale jednoduchá úloha lineární algebry a víme, že jde o nalezení kolmého průmětu vektoru Y do podprostoru $\langle X\rangle \subset \mathbb{R}^n$ generovaném sloupci matice X. Minimalizujeme přitom funkci

$$\|Y - X\beta\|^2 = \sum_{i=1}^{n} \Big(Y_i - \sum_{j=1}^{k} x_{ij}\beta_j\Big)^2.$$

Zvolme libovolnou ortonormální bázi vektorového podprostoru $\langle X\rangle$ a napišme ji do sloupců matice P. Pro jakoukoliv takovou volbu báze bude kolmý průmět realizován pomocí násobení maticí PP^T. Zobrazení dané touto maticí je přitom na podprostoru $\langle X\rangle$ identické, tj. dostáváme

$$\hat{Y} = PP^TY = PP^T(X\beta + \sigma Z) = X\beta + \sigma PP^TZ.$$

Matice PP^T je pozitivně semidefinitní. Nyní doplníme bázi ze sloupců v P na ortonormální bázi celého $\mathbb{R}^n$, tj. vytvoříme matici $Q = (P\ R)$ vepsáním nově přidaných vektorů báze do matice R s $n - k$ sloupci a n řádky. Označme si dále $V = P^TZ$ a $U = R^TZ$ náhodné vektory s k a $n - k$ komponentami. Budou na sebe vzájemně kolmé a jejich součtem v $\mathbb{R}^n$ dostaneme vektor $(V^TU^T)^T = Q^TZ$.

Evidentně tedy (viz odstavec 9.46) mají oba vektory V a U mnohoměrné normální rozdělení s nulovou střední hodnotou a jednotkovou kovarianční maticí. Náhodný vektor Y jsme rozložili na součet konstantního vektoru $X\beta$ a dvou kolmých projekcí

$$Y = X\beta + \sigma PV + \sigma RU$$

a hledaný kolmý průmět je součet prvních dvou sčítanců. V odstavci 9.46 jsme také pro takové náhodné vektory odvodili jejich rozdělení.

Velikost $\|Y - \hat{Y}\|^2$ nazýváme *reziduální součet čtverců*, zpravidla se značí *RSS*. Definujeme také *reziduální rozptyl* jako

$$S^2 = \frac{\|Y - Xb\|^2}{n-k}.$$

Připomeňme, že $\hat{Y} = Xb$ a že, díky našemu předpokladu o maximální hodnosti X, je matice X^TX invertibilní. Můžeme proto rovnou spočíst $b = (X^TX)^{-1}X^T\hat{Y}$. Zároveň ale víme, že

9.92. Ukažte, že v případě Bernoulliho procesu popsaného náhodnou veličinou $X \sim \mathrm{Bi}(n,\theta)$ a apriorní pravděpodobnosti náhodné veličiny θ s beta rozdělením, má i aposteriorní pravděpodobnost opět beta rozdělení s vhodnými parametry závislými na výsledku pokusu. Jaká bude aposteriorní střední hodnota veličiny θ (tj. bayesovský bodový odhad této náhodné veličiny)?

Řešení. Jak je zdůvodněno v odstavci 9.53 teoretického sloupce, bude aposteriorní hustota pravděpodobnosti, až na násobek vhodnou konstantou, dána jako součin apriorní hustoty pravděpodobnosti

$$g(\theta) = \frac{1}{\mathrm{B}(a,b)}\theta^{a-1}(1-\theta)^{b-1}$$

a pravděpodobnosti sledované veličiny X za podmínky, že nastala hodnota θ. Dostáváme tedy za předpokladu, že v Bernoulliho procesu nastalo k zdarů, aposteriorní hustotu (použitý znak místo rovnosti značí „proporcionální“)

$$\begin{aligned} g(\theta|X=k) &\propto P(X=k|\theta)g(\theta) \propto \\ &\propto \theta^k(1-\theta)^{n-k}\theta^{a-1}(1-\theta)^{b-1} = \\ &= \theta^{a+k-1}(1-\theta)^{b+n-k-1}. \end{aligned}$$

Dostali jsme tedy, až na konstantu, kterou nemusíme vůbec vyčíslovat, skutečně hustotu aposteriorního rozdělení pro veličinu θ s rozdělením $\mathrm{B}(a+k, b+n-k)$.

Její aposteriorní střední hodnota je

$$\hat\theta = \frac{a+k}{a+b+n}.$$

Pro n a k jdoucí do nekonečna, tak aby $k/n \to p$, bude i pro náš aposteriorní odhad platit $\hat\theta \to p$. Je tedy vidět, že při velkých hodnotách n a k bude převažovat pozorovaný podíl úspěšných pokusů nad apriorním předpokladem. Nicméně pro menší hodnoty je apriorní předpoklad naopak velice významný. □

9.93. Máme data o nehodovosti $N = 20$ řidičů za posledních $n = 10$ let (k–tá položka označuje počet roků, ve kterých došlo k nehodě u k–tého řidiče):

$$0, 0, 2, 0, 0, 2, 2, 0, 6, 4, 3, 1, 1, 1, 0, 0, 5, 1, 1, 0.$$

Předpokládáme, že pravděpodobnosti nehod u jednotlivých řidičů jsou konstanty p_j, $j = 1, \ldots, N$.

Odhadněte pro každého řidiče pravděpodobnost, že bude mít nehodu v následujícím roce (např. pro určení jeho individuálního pojistného). [2]

[2] Tento příklad je převzat z příspěvku M. Friesl, Bayesovské odhady v některých modelech, publikováno v: Analýza dat 2004/II (K. Kupka, ed.), Trilobyte Statistical Software, Pardubice, 2005, pp. 21-33.

$X^T(Y-\hat Y) = \sigma X^T(RU) = 0$, protože jsou sloupce v X a R po dvou ortogonální. Platí proto ve skutečnosti také

$$b = (X^TX)^{-1}X^TY. \tag{9.18}$$

Ještě můžeme lépe využít zvolenou matici P. Protože její sloupce generují tentýž podprostor jako sloupce X, jistě existuje čtvercová matice T taková, že $X = PT$ (její sloupce jsou koeficienty lineárních kombinací, které vyjadřují sloupce X pomocí báze v P). Nyní již jen dosadíme (použijeme přitom, že P^TP je jednotková matice a T je invertibilní):

$$\begin{aligned} b &= (T^TP^TPT)^{-1}T^TP^TY = \\ &= T^{-1}(T^T)^{-1}T^TP^T(PT\beta+\sigma Z) = \\ &= \beta + \sigma T^{-1}V. \end{aligned}$$

Tím jsme z velké části již odvodili vlastnosti lineárního modelu:

Věta. *Uvažujme lineární model $Y = X\beta + \sigma Z$.*
(1) Pro odhad $\hat Y$ platí

$$\hat Y = X\beta + \sigma PV, \quad \hat Y \sim \mathrm{N}(X\beta, \sigma^2PP^T).$$

(2) Reziduální součet čtverců RSS a normovaný čtverec velikosti rezidua mají rozdělení:

$$Y - \hat Y \sim \mathrm{N}(0, \sigma^2RR^T), \quad \|Y-Y\|^2/\sigma^2 \sim \chi^2_{n-k}.$$

(3) Náhodná veličina $b = \beta + \sigma T^{-1}V$ má rozdělení

$$b \sim \mathrm{N}(\beta, \sigma^2(X^TX)^{-1}).$$

(4) Pro reziduální rozptyl platí $(n-k)S^2/\sigma^2 \sim \chi^2_{n-k}$.
(5) Střední hodnota reziduálního rozptylu je $\mathrm{E}\,S^2 = \sigma^2$.
(6) Veličiny b a S^2 jsou nezávislé.

Důkaz. Tvar i rozdělení $\hat Y$ jsme již odvodili. Odtud je ale již zřejmé, že $Y - \hat Y = \sigma RU$ a tím máme ověřené i druhé tvrzení. Dále máme

$$\|Y-Y\|^2/\sigma^2 = \|RU\|^2 = \|U\|^2,$$

kde poslední rovnost plyne z toho, že v naší konstrukci je U vektor souřadnic průmětu Z do komplementu $\langle X\rangle$ a RU je tímto průmětem. Velikost vektoru je přitom dána právě jako součet kvadrátů souřadnic v libovolné ortonormální bázi.

Je proto náhodný vektor $\|Y-Y\|^2/\sigma^2$ součtem $(n-k)$ kvadrátů náhodných veličin s rozdělením $\mathrm{N}(0,1)$, tedy jde o rozdělení χ^2_{n-k} a dokázali jsme zbytek (2).

Další tvrzení vyplývá přímo z našich definic a výpočtů, zbývá jen odhadnout varianční matici pro b. Z obecných vlastností víme, že má vyjít matice $T^{-1}(T^T)^{-1}$. To je ale stejná matice jako $(X^TX)^{-1} = ((PT)^T(PT))^{-1}$.

Tvrzení z (4) je jen přepsáním informace z (2) a další tvrzení plyne z toho, že střední hodnota rozdělení χ^2 je rovna počtu stupňů volnosti.

Konečně, nezávislost veličin b a S je důsledkem toho že první je funkcí vektoru V, zatímco druhá je funkcí vektoru U a tyto vektory jsou nezávislé, protože vznikly jako dvě komplementární části z ortogonální transformace vektoru Z. □

V praktických úlohách někdy testujeme hypotézu, zda pro odhadu středních hodnot nestačí méně parametrů. Říkáme, že náhodný vektor Y splňuje *podmodel*, když platí zároveň $Y = X\beta + \sigma Z$ a

Řešení. Zavedeme si náhodné veličiny X_{ij} s hodnotami 0, když i-tý řidič v j-tém roce neměl žádnou nehodu, a hodnotami 1 pokud nehodu měl. Jednotlivé roky považujeme za nezávislé, můžeme proto předpokládat, že náhodné veličiny $S_j = \sum_{i=1}^{n} X_{ji}$ udávající počet nehod za všech $n = 10$ let mají rozdělení $\mathrm{Bi}(n, p_j)$.

Samozřejmě bychom mohli odhadnout pravděpodobnosti pro všechny řidiče společně, tj. pomocí aritmetického průměru

$$\hat{p} = \frac{1}{N}\sum_{j=1}^{n} S_j \frac{1}{n} = \frac{1}{20}\frac{29}{10} = 0{,}145.$$

Když ale uvážíme homogennost rozdělení veličin X_j, těžko je lze považovat za shodné, proto bude takovýto odhad jistě zavádějící.

Opačný extrém, tj. zcela nezávislý a individuální odhad

$$\hat{p}_j = \frac{1}{n} S_j$$

je samozřejmě také nevhodný, protože jistě nechceme předepisovat nulové pojistné, dokud nedojde k první nehodě.

Jako realistický se jeví postup, ve kterém využijeme stejný předpoklad apriorního rozdělení pravděpodobnosti p_j nehodovosti u jednotlivých řidičů. V praxi se zpravidla používá model s Poissonovým rozdělením $\mathrm{Po}(\lambda_j)$ u j-tého řidiče s dalšími předpoklady o rozdělení parametru λ mezi řidiči. Docela dobře (a hlavně jednoduše) můžeme také předpokládat, že v našem případě půjde o rozdělení $p_j \sim \beta(a, b)$ s vhodnými parametry a, b, které by měly odrážet kumulované výsledky všech řidičů. Pojďme tedy touto cestou.

Z předchozího příkladu pak víme, že aposteriorní rozdělení pravděpodobností bude $(p_j | S_j = k) = \beta(a + k, b + n - k)$, takže příslušná střední hodnota bude

$$\hat{p}_j^b = \frac{a + k}{a + b + n}.$$

Srovnejme si tento odhad s výše uvedeným společným odhadem $\hat{p}$ a individuálním $\hat{p}_j$. Zaveďme si k tomu hodnoty $p_0 = \frac{a}{a+b}$, tj. střední hodnotu apriorního společného rozdělení pro všechny řidiče, a $n_0 = a + b$. Dostáváme

$$\hat{p}_j^b = \frac{(a+b)a}{(a+b+n)(a+b)} + \frac{nk}{(a+b+n)n} = \frac{n_0}{n_0+n} p_0 + \frac{n}{n_0+n}\hat{p}_j,$$

což je lineární kombinace střední hodnoty p_0 a individuálního odhadu $\hat{p}_j$.

Zbývá nám tedy už jen smysluplně odhadnout neznámé parametry a, b. Víme přitom

$$\mathrm{E}\, X_{ji} = \mathrm{E}\,\mathrm{E}(X_{ji}\,|\,p) = \mathrm{E}\, p = p_0$$
$$\frac{\mathrm{E}\,\mathrm{var}(X_{ji}\,|\,p)}{\mathrm{var}\,\mathrm{E}(X_{ji}\,|\,p)} = \frac{\mathrm{E}(p(1-p))}{\mathrm{var}\, p} = a + b = n_0$$

$$Y = X^0 \beta^0 + \sigma Z,$$

kde X^0 má jen $q < k$ sloupců a předpokládáme, že sloupce X^0 generují podprostor v $\langle X \rangle$, tj. jsou všechny lineárními kombinacemi sloupců v X.

Nyní můžeme zopakovat předchozí konstrukci a zvolit přitom matici P tak, aby jejích prvních q vektorů generovalo $\langle X^0 \rangle$. Celé P tak bude mít tvar $(P^0\ P^1)$ a stejně tak se rozpadne i vektor V

$$V = \begin{pmatrix} V^0 \\ V^1 \end{pmatrix} = \begin{pmatrix} (P^0)^T Z \\ (P^1)^T Z \end{pmatrix}.$$

Dostáváme tak jemnější rozklad vektorů a jejich velikostí a příslušných reziduí

$$\hat{Y}^0 = P^0 (P^0)^T Y = X^0 \beta^0 + \sigma P^0 V^0$$
$$Y - \hat{Y}^0 = \sigma P^1 V^1 + \sigma R U$$
$$\|Y - \hat{Y}^0\|^2 = \sigma^2 \|V^1\|^2 + \sigma^2 \|U\|^2 (\mathrm{RSS}^0 - \mathrm{RSS})/\sigma^2 = \|V^1\|^2.$$

Normovaný rozdíl reziduí má tedy rozdělení χ^2_{k-q}. Odtud okamžitě vyplývá, že statistika F zadaná jako relativní rozdíl reziduí má Fischerovo-Snedecorovo rozdělení

$$F = \frac{(\mathrm{RSS}^0 - \mathrm{RSS})/(k-q)}{\mathrm{RSS}\,/(n-k)} \sim \mathrm{F}_{k-q,n-k}\ .$$

Často v praktických situacích skutečně neznáme parametr σ a nahrazujeme jej jeho odhadem S^2. Místo jednotlivých složek $b_j \sim \mathrm{N}(\beta_j, \sigma^2 c_{jj})$ náhodného vektoru b, kde c_{jj} jsou diagonální prvky v matici $C = (X^T X)^{-1}$, pak pracujeme se statistikami

$$T_j = \frac{b_j - \beta_j}{S\sqrt{c_{jj}}} \sim \mathrm{t}_{n-k}\ .$$

Tyto veličiny již samozřejmě nemusí být nezávislé.

V případě, že bychom nepředpokládali plnou hodnost matice X, používali bychom v obdobných úvahách místo matice $C = (X^T X)^{-1}$ matici pseudoinverzní.

9.57. Příklady testů. Jako ilustraci velmi stručně zmíníme

několik příkladů použití lineárních modelů v nejjednodušších typech testů. Úplně nejjednodušší je to v případě jediného výběru, kdy testujeme, zda jediný parametr β je roven dané hodnotě β_0.

Pro tento případ můžeme zvolit matici X s jediným sloupcem plným jedniček. Výraz

$$Y = X\beta + \sigma Z$$

tedy značí, že jednotlivé komponenty v Y jsou nezávislé veličiny $Y_i \sim N\beta, \sigma^2$, jde tedy obvyklý náhodný výběr rozsahu n z normálního rozdělení. Z našich obecných úvah okamžitě vidíme odhad

$$b = (X^T X)^{-1} X^T Y = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$$

$$S^2 = \frac{1}{n-1}\|Y - X\bar{Y}\|^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

což jsou právě výběrový průměr a rozptyl, se kterými jsme již počítali.

Zajímavá je v tomto kontextu hlavně statistika

$$T = \frac{\bar{Y} - \beta_0}{S}\sqrt{n}$$

a přitom veličiny na levých stranách můžeme přímo odhadnout.

$$\mathrm{E}\,X_{ij} = \mathrm{E}\,\mathrm{E}(X_{ji}\,|\,p) \simeq \frac{1}{N}\sum_{j=1}^{N}\hat{p}_j$$

$$\mathrm{E}\,\mathrm{var}(X_{ji}\,|\,p) \simeq \frac{1}{N}\sum_{j=1}^{N}\Big(\frac{n}{n-1}\hat{p}_j(1-\hat{p}_j)\Big)$$

$$\mathrm{var}\,\mathrm{E}(X_{ji}\,|\,p) \simeq s^2_{\hat{p}_j} - \frac{1}{nN}\sum_{j=1}^{N}\Big(\frac{n}{n-1}\hat{p}_j(1-\hat{p}_j)\Big),$$

kde $s^2_{\hat{p}_j}$ označuje výběrový rozptyl mezi individuálními odhady (čtenář si může promyslet, že odečtením posledního výrazu vpravo zajišťujeme, aby i poslední odhad byl nestranný).

Protože pro uvedená data takto dostáváme $n_0 \simeq 3{,}8643$ a $p_0 \simeq 0{,}1450$, vyjde nám bayesovský odhad individuální pravděpodobnosti nehod

$$\hat{p}^b_j = 0{,}154 \cdot 0{,}145 + 0{,}846 \cdot \hat{p}_j.$$

Jde tedy o kombinaci spolehlivého odhadu $\hat{p} = 0{,}145$ kolektivní pravděpodobnosti p_0 s individuálním (frekvenčním) odhadem $\hat{p}_j$, který je pořízen z malého počtu pozorování $n = 10$ u jediného řidiče. □

L. Zpracování vícerozměrných dat

Někdy potřebujeme zpracovat vícerozměrná data: u každého z n objektů určujeme p znaků. Například můžeme zkoumat známky různých žáků z různých předmětů.

9.94. Ve svých pokusech pozoroval J.G.Mendel 10 rostlin hrachu a na každé z nich počet žlutých a zelených semen. Výsledky pokusu jsou shrnuty v následující tabulce:

číslo rostliny	1	2	3	4	5	6	7	8	9	10
počet žlutých	25	32	14	70	24	20	32	44	50	44
počet zelených	11	7	5	27	13	6	13	9	14	18
celkem	36	39	19	97	37	26	45	53	64	62

Z genetických modelů vyplývá, že pravděpodobnost výskytu žlutého semene by měla být 0,75 a zeleného 0,25. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že se výsledky Mendelových pokusů shodují s modelem.

Řešení. Hypotézu budeme testovat *testem dobré shody*. Použijeme statistiku

$$K = \sum_{j=1}^{r}\frac{(n_j - np_j)}{np_j},$$

Testování hypotézy $\beta = \beta_0$ se nazývá *jednovýběrový t-test*. Na hladině α hypotézu zamítáme, když je $|T| \geq t_{n-1}(\alpha)$.

Obdobné jednoduché využití obecného modelu se nazývá *párový t-test*. Je vhodný na případy, kdy testujeme dvojice náhodných vektorů $W_1 = (W_{i1})$ a $W_2 = (W_{i2})$, o rozdílech jejichž komponent $Y_i = W_{i1} - W_{i2}$ víme, že mají rozdělení $\mathrm{N}(\beta, \sigma^2)$. Potřebujeme navíc, aby byly veličiny Y_i nezávislé (což neříká, že musí být nezávislé jednotlivé dvojice W_{i1} a W_{i2}!). Můžeme si v kontextu našeho ilustračního příkladu z 9.50 představit třeba hodnocení dvou různých vyučujících týmž studentem.

Testujeme hypotézu, že pro všechna i je $\mathrm{E}\,W_{i1} = \mathrm{E}\,W_{i2}$. Je zjevné, že $Y = W_1 - W_2$ bude splňovat Používáme tedy statistiku

$$T = \frac{\bar{W}_1 - \bar{W}_2}{S}\sqrt{n}.$$

Zmíníme ještě jeden jednoduchý příklad s více parametry. Půjde o klasický případ *regresní přímky*.

Předpokládáme, že veličiny Y_i, $i = 1, \ldots, n$ mají rozdělení $\mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2)$, kde x_i jsou dané konstanty. Zkoumáme tedy co nejlepší aproximaci

$$Y_i = b_0 + b_1 x_i$$

a matice X příslušného lineárního modelu je

$$X^T = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \end{pmatrix}.$$

Dosazením do obecných vztahů snadno spočteme odhad

$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} n & \bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^n x_i Y_i \end{pmatrix} =$$
$$= \left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^{-1} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^n x_i Y_i. \end{pmatrix}$$

Odtud už po drobné úpravě vychází

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

a konečně spočteme $b_0 = \bar{Y} - b_1\bar{x}$. Z výpočtu je přitom vidět, že

$$\mathrm{var}\,b_1 = \sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2.$$

Pro testování hypotézy, zda střední hodnota veličiny Y nezávisí na x, tj. H_0 je tvaru $\beta_1 = 0$, můžeme tedy použít statistiku

$$T = \frac{b_1}{S}\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^{1/2} \sim \mathrm{t}_{n-2}\,.$$

Úplně obdobně vypadá statistická analýza vícenásobné regrese, kde máme několik sad hodnot x_{ij} a vyhodnocujeme statistickou relevanci aproximace

$$Y_i = b_0 + b_1 x_{1i} + \cdots + b_k x_{ki}.$$

Jednotlivé statistiky T_j zde umožňují t-test závislosti regrese na jednotlivých parametrech. Softwarové balíčky zpravidla uvádí také parametr vyjadřující, jak dobře jsou celkově hodnoty Y_i aproximovány. Říkává se mu koeficient determinace

$$R^2 = 1 - \frac{\mathrm{RSS}}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

kde r je počet třídících intervalů (měření; v našem případě $r = 10$), n_j je skutečně naměřená četnost znaku ve zvoleném třídícím intervalu (budeme počítat množství žlutých semen), p_j očekávaná četnost (podle předpokládaného rozložení), v našem případě $p_j = 0{,}75$, $j = 1, \ldots, 10$. Pokud by se výsledky pokusu skutečně řídily naším modelem, pak by $K \approx \chi^2(r - 1 - p)$, kde p je počet odhadovaných parametrů v předpokládaném rozložení pravděpodobnosti. V našem případě je to obzvláště jednoduché, neboť náš model žádné neznámé parametry neobsahuje, je tedy $p = 0$ (parametry se mohou vyskytnout, pokud například předpokládáme, že rozložení pravděpodobnosti v našem pokusu bude normální, ovšem s neznámým rozptylem a střední hodnotou; potom by $p = 2$). Bude tedy $K \approx \chi^2(9)$. Statistika se doporučuje používat, pokud je očekávaná četnost znaku v každém z třídících intervalů alespoň pět.

Zapišme data do tabulky:

j	n_j	p_j	np_j	$\frac{(n_j - np_j)^2}{np_j}$
1	25	0,75	27	0,148148
2	32	0,75	29, 25	0,258547
⋮	⋮	⋮	⋮	⋮
10	44	0,75	46,5	0,134409

Hodnota statistiky K pro daná data je

$$K = 0{,}148148 + 0{,}258547 + \cdots + 0{,}134409 = 1{,}797495.$$

Tato hodnota je menší než $\chi^2_{095}(9) = 16{,}9$, nulovou hypotézu na hladině významnosti 0,05 tedy nezamítáme (nevylučujeme tedy, že známý model dědičnosti skutečně platí).

□

9.58. V praktických situacích se velmi často setkáváme s problémy, kdy jsou buď rozdělení základních statistických souborů úplně neznámá nebo jsou v modelu předpokládané chyby a odchylky s nenulovou střední hodnotou a jiným než normálním rozdělením. V těchto situacích je využití klasické frekvenční statistiky buď velmi obtížné nebo zcela nemožné.

Existují ale přístupy, jak pracovat přímo nad výběrovým souborem a odvozovat statistiky bodových či intervalových odhadů nebo pravděpodobnostní úsudky v obdobě k předchozím bodům, včetně vyčíslování standardních chyb.

Jedním ze zásadních průkopnických článků v tomto směru byla již v roce 1981 publikovaná stručná práce Bradleyho Efrona ze Stanfordu *Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods*.[4]. Klíčová slova v tomto článku jsou: Balanced repeated replications; Bootstrap; Delta method; Half-sampling; Jackknife; Infinitesimal jackknife; Influence function.

Nemáme tu prostor pro podrobnější rozbor těchto technik, které jsou základem neparametrických metod v současných softwarových statistických nástrojích. Pro ilustraci jen stručně zmiňme postup v metodě *bootstrap*. Softwarovými prostředky v tomto případě z daného výběrového souboru vytváříme nové a nové výběrové soubory stejného rozsahu (přičemž skutečně provádíme výběr s vrácením vybrané jednotky do základního souboru) a pro každý z nich sledujeme potřebné statistiky (výběrový průměr, rozptyl apod.). Po velkém počtu opakování tohoto postupu tak získáme soubor, který považujeme za relevantní přiblíženi pravděpodobnostního rozložení zkoumané statistiky. Charakteristiky tohoto souboru považujeme za dobré přiblížení charakteristik zkoumané statistiky při bodových či intervalových odhadech, analýze rozptylu apod.

[4] Biometrika (1981), 68, 3, pp. 589-99

Řešení cvičení

9.21. $\frac{3}{5} \cdot \frac{2}{3} + \frac{2}{5} \cdot 1 = \frac{4}{5}$.

9.41. Jednoduše $a = \frac{3}{8}$. Distribuční funkce náhodné veličiny X je tedy $F_X(t) = \frac{1}{8}t^3$ pro $t \in (0, 2)$, pro menší t je tato funkce nulová, pro větší rovna 1. Označme $Z = X^3$ náhodnou veličinu označující objem krychle. Ten je v intervalu $(0, 8)$, pro $t \in (0, 8)$ a distribuční funkci F_Z náhodné veličiny Z tedy můžeme psát $F_Z(t) = P[Z < t] = P[X^3 < t] = P[X < \sqrt[3]{t}] = F_X(\sqrt[3]{t}) = \frac{1}{8}t$, hustota pravděpodobnosti je pak $f_Z(t) = \frac{1}{8}$ na intervalu $(0, 8)$, jinak nula, jedná se tedy o rovnoměrné rozdělení pravděpodobnosti na daném intervalu, střední hodnota je tudíž 4.

9.55. $EU = 1 \cdot 0{,}6 + 2 \cdot 0{,}4 = 1{,}4$, $EU^2 = 0{,}4 + 4 \cdot 0{,}4 = 2{,}2$ $EV = 0{,}3 + 0{,}6 + 1{,}2 = 2{,}1$, $EV^2 = 0{,}3 + + 1, 2 + 3{,}6 = 5{,}1$, $E(UV) = 2{,}8$, $\mathrm{var}(U) = 2{,}2 - 1{,}4^2 = 2{,}2 - 1{,}96 = 0{,}24$, $\mathrm{var}(V) = 5{,}1 - 4{,}41 = 0{,}69$, $\mathrm{cov}(UV) = 2{,}8 - 1{,}4 \cdot 2{,}1 = -0{,}14$, $\rho_{U,V} = \frac{-0{,}14}{\sqrt{0{,}24 \cdot 0{,}69}} \doteq 0, 39$.

9.56. $EX = 1/3$, $\mathrm{var}^2\, X = 4/45$.

9.57.

$$\rho_{X,Y} = -1.$$

9.58. $\varrho_{U,V} \doteq -0, 421$.